Spectral Methods and Computational Trade-offs in High-dimensional Statistical Inference

Tengyao Wang

St John's College

University of Cambridge

A thesis submitted for the degree of

Doctor of Philosophy

July 2016

Abstract

Spectral methods have become increasingly popular in designing fast algorithms for modern high-dimensional datasets. This thesis looks at several problems in which spectral methods play a central role. In some cases, we also show that such procedures have essentially the best performance among all randomised polynomial time algorithms by exhibiting statistical and computational trade-offs in those problems.

In the first chapter, we prove a useful variant of the well-known Davis–Kahan theorem, which is a spectral perturbation result that allows us to bound of the distance between population eigenspaces and their sample versions.

We then propose a semi-definite programming algorithm for the sparse principal component analysis (PCA) problem, and analyse its theoretical performance using the perturbation bounds we derived earlier. It turns out that the parameter regime in which our estimator is consistent is strictly smaller than the consistency regime of a minimax optimal (yet computationally intractable) estimator. We show through reduction from a well-known hard problem in computational complexity theory that the difference in consistency regimes is unavoidable for any randomised polynomial time estimator, hence revealing subtle statistical and computational trade-offs in this problem.

Such computational trade-offs also exist in the problem of restricted isometry certification. Certifiers for restricted isometry properties can be used to construct design matrices for sparse linear regression problems. Similar to the sparse PCA problem, we show that there is also an intrinsic gap between the class of matrices certifiable using unrestricted algorithms and using polynomial time algorithms.

Finally, we consider the problem of high-dimensional changepoint estimation, where we estimate the time of change in the mean of a high-dimensional time series with piecewise constant mean structure. Motivated by real world applications, we assume that changes only occur in a sparse subset of all coordinates. We apply a variant of the semi-definite programming algorithm in sparse PCA to aggregate the signals across different coordinates in a near optimal way so as to estimate the changepoint location as accurately as possible. Our statistical procedure shows superior performance compared to existing methods in this problem.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specically indicated in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution.

Chapter 1 is joint work with Yi Yu (University of Cambridge) and Richard Samworth, and has appeared in *Biometrika* as Yu, Wang and Samworth (2015). Chapter 2 is joint work with Quentin Berthet (University of Cambridge) and Richard Samworth, and is to appear in *Annals of Statistics* as Wang, Berthet and Samworth (2016). Chapter 3 is joint work with Quentin Berthet and Yaniv Plan (University of British Columbia) and Chapter 4 is joint work with Richard Samworth. Both Chapter 3 and Chapter 4 have been submitted for publication (Wang, Berthet and Plan, 2016; Wang and Samworth, 2016a).

Tengyao Wang
Cambridge, July 2016

Acknowledgements

I owe my deepest gratitude to my supervisor Professor Richard Samworth. Without his encouragement and guidance this PhD dissertation would not have been possible. His Part III course first brought me into the exciting area of high-dimensional statistics and persuaded me into pursuing academic research in this area. I am very lucky to have collaborated with him on several projects during my PhD and I learned a tremendous amount while working with him. I would also like to thank my other collaborators, Quentin Berthet, Yi Yu and Yaniv Plan, for introducing me to interesting problems and working together to solve them. I thank all the people at Statslab for creating a friendly and vibrant environment here. In particular, I would like to thank Tim Cannings, Tom Berrett, Yining Chen, Arlene Kim, Danning Li, Rajen Shah, Jenny Wadsworth and Yi Yu for their advice and helpful discussions.

I gratefully acknowledge the funding sources that made my PhD work possible. I was funded by a Benefactors' Scholarship from St John's College, Cambridge and Cambridge Overseas Trust.

Lastly, I would like to thank my family for all their love and encouragement. For my parents who raised me with a love of mathematics and supported me in all my pursuits. And most of all for my loving wife Chenqu for her encouragement and support along the way. Thank you.

Contents

Chapter 1

A useful variant of the Davis–Kahan theorem for statisticians

1.1 Introduction

Many statistical procedures rely on the eigendecomposition of a matrix. Examples include principal components analysis and its cousin sparse principal components analysis (Zou, Hastie and Tibshirani, 2006), factor analysis, high-dimensional covariance matrix estimation (Fan, Liao and Mincheva, 2013) and spectral clustering for community detection with network data (Donath and Hoffman, 1973). In these and most other related statistical applications, the matrix involved is real and symmetric, e.g. a covariance or correlation matrix, or a graph Laplacian or adjacency matrix in the case of spectral clustering.

In the theoretical analysis of such methods, it is frequently desirable to be able to argue that if a sample version of this matrix is close to its population counterpart, and provided certain relevant eigenvalues are well-separated in a sense to be made precise below, then a population eigenvector should be well approximated by a corresponding sample eigenvector. A quantitative version of such a result is provided by the Davis–Kahan $\sin\theta$ theorem (Davis and Kahan, 1970). This is a deep theorem from operator theory, involving operators acting on Hilbert spaces, though as remarked by Stewart and Sun (1990), its 'content more than justifies its impenetrability'. In statistical applications, we typically do not require this full generality; in Theorem 1.1 below, we state a version in a form typically used in the statistical literature (e.g. von Luxburg, 2007; Rohe, Chatterjee and Yu, 2011). Since the theorem allows for the possibility that more than one eigenvector is of interest, we need to define a notion of distance between subspaces spanned by two sets of vectors. This can be done through the idea of principal angles: if $V, \hat{V} \in \mathbb{R}^{p\times d}$ both have orthonormal columns, then the vector of d principal angles between their column spaces is $(\cos^{-1}\sigma_1, \ldots, \cos^{-1}\sigma_d)^\top$, where $\sigma_1 \geq \cdots \geq \sigma_d$ are the singular values of $\hat{V}^\top V$. Thus, principal angles between subspaces can be considered as a natural generalisation of the acute angle between two vectors. We let $\Theta(\hat{V}, V)$ denote the $d \times d$ diagonal matrix whose jth diagonal entry is the jth principal angle, and let $\sin\Theta(\hat{V}, V)$ be defined entrywise. A convenient way to measure the

distance between the column spaces of V and $\hat{V}$ is via $\|\sin\Theta(\hat{V}, V)\|_{\mathrm{F}}$, where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm of a matrix.

Theorem 1.1 (Davis–Kahan $\sin\theta$ theorem)**.** *Let $\Sigma, \hat{\Sigma} \in \mathbb{R}^{p\times p}$ be symmetric, with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$ respectively. Fix $1 \leq r \leq s \leq p$, let $d := s - r + 1$, and let $V = (v_r, v_{r+1}, \ldots, v_s) \in \mathbb{R}^{p\times d}$ and $\hat{V} = (\hat{v}_r, \hat{v}_{r+1}, \ldots, \hat{v}_s) \in \mathbb{R}^{p\times d}$ have orthonormal columns satisfying $\Sigma v_j = \lambda_j v_j$ and $\hat{\Sigma}\hat{v}_j = \hat{\lambda}_j \hat{v}_j$ for $j = r, r+1, \ldots, s$. Write $\delta = \inf\{|\hat{\lambda} - \lambda| : \lambda \in [\lambda_s, \lambda_r], \hat{\lambda} \in (-\infty, \hat{\lambda}_{s+1}] \cup [\hat{\lambda}_{r-1}, \infty)\}$, where we define $\hat{\lambda}_0 := \infty$ and $\hat{\lambda}_{p+1} := -\infty$, and assume that $\delta > 0$. Then*

$$\|\sin\Theta(\hat{V}, V)\|_{\mathrm{F}} \leq \frac{\|\hat{\Sigma} - \Sigma\|_{\mathrm{F}}}{\delta}. \tag{1.1}$$

Theorem 1.1 is an immediate consequence of Theorem V.3.6 of Stewart and Sun (1990). Despite the attractions of this bound, an obvious difficulty for statisticians is that we may have $\delta = 0$ for a particular realisation of $\hat{\Sigma}$, even when the population eigenvalues are well-separated. As a toy example to illustrate this point, suppose that $\Sigma = \mathrm{diag}(50, 40, 30, 20, 10)$ and $\hat{\Sigma} = \mathrm{diag}(54, 37, 32, 23, 21)$. If we are interested in the eigenspaces spanned by the eigenvectors corresponding to the second, third and fourth largest eigenvalues, so $r = 2$ and $s = 4$, then Theorem 1.1 above cannot be applied, because $\delta = 0$.

Ignoring this issue for the moment, we remark that both occurrences of the Frobenius norm in (1.1) can be replaced with the operator norm $\|\cdot\|_{\mathrm{op}}$, or any other orthogonally invariant norm. Frequently in applications, we have $r = s = j$, say, in which case we can conclude that

$$\sin\Theta(\hat{v}_j, v_j) \leq \frac{\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}}{\min(|\hat{\lambda}_{j-1} - \lambda_j|, |\hat{\lambda}_{j+1} - \lambda_j|)}.$$

Since we may reverse the sign of $\hat{v}_j$ if necessary, there is a choice of orientation of $\hat{v}_j$ for which $\hat{v}_j^{\top} v_j \geq 0$. For this choice, we can also deduce that $\|\hat{v}_j - v_j\| \leq 2^{1/2}\sin\Theta(\hat{v}_j, v_j)$, where $\|\cdot\|$ denotes the Euclidean norm.

Theorem 1.1 is typically used to show that $\hat{v}_j$ is close to v_j as follows: first, we argue that $\hat{\Sigma}$ is close to Σ. This is often straightforward; for instance, when Σ is a population covariance matrix, it may be that $\hat{\Sigma}$ is just an empirical average of independent and identically distributed random matrices; cf. Section 1.3. Then we argue, e.g. using Weyl's inequality (Weyl, 1912; Stewart and Sun, 1990), that with high probability, $|\hat{\lambda}_{j-1} - \lambda_j| \geq (\lambda_{j-1} - \lambda_j)/2$ and $|\hat{\lambda}_{j+1} - \lambda_j| \geq (\lambda_j - \lambda_{j+1})/2$, so on these events $\|\hat{v}_j - v_j\|$ is small provided we are also willing to assume an eigenvalue separation, or eigen-gap, condition on the population eigenvalues.

The main purpose of our work in this chapter, in Theorem 1.2 in Section 1.2 below, is to give a variant of the Davis–Kahan $\sin\theta$ theorem that has two advantages for statisticians. First, the only eigen-gap condition is on the population eigenvalues, in contrast to the definition of δ in Theorem 1.1 above. Similarly, only population eigenvalues appear in the denominator of the bounds. This means there is no need for the statistician to worry about the event where $|\hat{\lambda}_{j-1} - \lambda_j|$ or $|\hat{\lambda}_{j+1} - \lambda_j|$ is small. Second, we show that the expression $\|\hat{\Sigma} - \Sigma\|_{\mathrm{F}}$ appearing in the numerator of the bound in (1.1) can be replaced with $\min(d^{1/2}\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}, \|\hat{\Sigma} - \Sigma\|_{\mathrm{F}})$. In Section 1.3, we give applications where our result could be used to allow authors to assume more natural conditions or to simplify proofs, and also give a detailed example to illustrate the potential improvements of our bounds. Our result is also used to provide theoretical control for the spectral methods proposed in Chapter 2 and Chapter 4. The recent result of Vu et al. (2013, Corollary 3.1) has some overlap

with our Theorem 1.2. We discuss the differences between our work and theirs shortly after the statement of Theorem 1.2.

Singular value decomposition, which may be regarded as a generalisation of eigendecomposition, but which exists even when a matrix is not square, also plays an important role in many modern algorithms in statistics and machine learning. Examples include matrix completion (Candès and Recht, 2009), robust principal components analysis (Candès et al., 2009) and motion analysis (Kukush, Markovsky and Van Huffel, 2002), among many others. Wedin (1972) provided the analogue of the Davis–Kahan $\sin\theta$ theorem for such general real matrices, working with singular vectors rather than eigenvectors, but with conditions and bounds that mix sample and population singular values. In Section 1.4, we extend the results of Section 1.2 to such settings; again our results depend only on a condition on the population singular values. Proofs are deferred to Section 1.5.

1.2 Main results

Theorem 1.2. *Let $\Sigma, \hat{\Sigma} \in \mathbb{R}^{p\times p}$ be symmetric, with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$ respectively. Fix $1 \leq r \leq s \leq p$ and assume that $\min(\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1}) > 0$, where we define $\lambda_0 := \infty$ and $\lambda_{p+1} := -\infty$. Let $d := s - r + 1$, and let $V = (v_r, v_{r+1}, \ldots, v_s) \in \mathbb{R}^{p\times d}$ and $\hat{V} = (\hat{v}_r, \hat{v}_{r+1}, \ldots, \hat{v}_s) \in \mathbb{R}^{p\times d}$ have orthonormal columns satisfying $\Sigma v_j = \lambda_j v_j$ and $\hat{\Sigma}\hat{v}_j = \hat{\lambda}_j \hat{v}_j$ for $j = r, r+1, \ldots, s$. Then*

$$\|\sin\Theta(\hat{V}, V)\|_{\mathrm{F}} \leq \frac{2\min(d^{1/2}\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}, \|\hat{\Sigma} - \Sigma\|_{\mathrm{F}})}{\min(\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1})}. \tag{1.2}$$

Moreover, there exists an orthogonal matrix $\hat{O} \in \mathbb{R}^{d\times d}$ such that

$$\|\hat{V}\hat{O} - V\|_{\mathrm{F}} \leq \frac{2^{3/2}\min(d^{1/2}\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}, \|\hat{\Sigma} - \Sigma\|_{\mathrm{F}})}{\min(\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1})}. \tag{1.3}$$

We remark that even though we have stated Theorem 1.2 for V and $\hat{V}$ being matrices of eigenvectors corresponding to blocks of consecutive eigenvalues, which is the most interesting case in statistical applications, the same result holds if we let V and $\hat{V}$ be eigenvectors corresponding to an arbitrary subset of the spectra of Σ and $\hat{\Sigma}$ respectively. More precisely, let $J \subseteq \{1, \ldots, p\}$, define $\lambda_J := \{\lambda_j : j \in J\}$, $\hat{\lambda}_J := \{\hat{\lambda}_j : j \in J\}$, and let $V, \hat{V} \in \mathbb{R}^{p\times d}$ have orthonormal columns that are respectively eigenvectors corresponding to λ_J and $\hat{\lambda}_J$. Then we have

$$\|\sin\Theta(\hat{V}, V)\|_{\mathrm{F}} \leq \frac{2\min(d^{1/2}\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}, \|\hat{\Sigma} - \Sigma\|_{\mathrm{F}})}{\min\{|\lambda_j - \lambda_{j'}| : j \in J, j' \notin J\}},$$

the proof which is almost exactly the same as the proof of Theorem 1.2 after changing $\min(\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1})$ to $\min\{|\lambda_j - \lambda_{j'}| : j \in J, j' \notin J\}$ when necessary.

As mentioned briefly in the introduction, apart from the fact that we only impose a population eigen-gap condition, the main difference between this result and that given in Theorem 1.1 is in the $\min(d^{1/2}\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}, \|\hat{\Sigma} - \Sigma\|_{\mathrm{F}})$ term in the numerator of the bounds. In fact, the original statement of the Davis–Kahan $\sin\theta$ theorem has a numerator of $\|V\Lambda - \hat{\Sigma}V\|_{\mathrm{F}}$ in our notation, where $\Lambda = \mathrm{diag}(\lambda_r, \lambda_{r+1}, \ldots, \lambda_s)$. However, in order to apply that theorem in practice, statisticians have bounded this expression by $\|\hat{\Sigma} - \Sigma\|_{\mathrm{F}}$, yielding the bound in Theorem 1.1. When p is large, though,

one would often anticipate that $\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}$, which is the ℓ_∞ norm of the vector of eigenvalues of $\hat{\Sigma} - \Sigma$, may well be much smaller than $\|\hat{\Sigma} - \Sigma\|_{\mathrm{F}}$, which is the ℓ_2 norm of this vector of eigenvalues. Thus when $d \ll p$, as will often be the case in practice, the minimum in the numerator may well be attained by the first term. It is immediately apparent from (1.8) and (1.9) in our proof that the smaller numerator $\|\hat{V}\Lambda - \Sigma\hat{V}\|_{\mathrm{F}}$ could also be used in our bound for $\|\sin\Theta(\hat{V}, V)\|_{\mathrm{F}}$ in Theorem 1.2, while $2^{1/2}\|\hat{V}\Lambda - \Sigma\hat{V}\|_{\mathrm{F}}$ could be used in our bound for $\|\hat{V}\hat{O} - V\|_{\mathrm{F}}$. Our reason for presenting the weaker bound in Theorem 1.2 is to aid direct applicability; see Section 1.3 for examples.

As mentioned in the introduction, Vu et al. (2013, Corollary 3.1) is similar in spirit to Theorem 1.2 above, and only involves a population eigen-gap condition, but there are some important differences. First, their result focuses on the eigenvectors corresponding to the top d eigenvalues, whereas ours applies to any set of d eigenvectors corresponding to a block of d consecutive eigenvalues, as in the original Davis–Kahan theorem. Their proof, which uses quite different techniques from ours, does not appear to generalise immediately to this setting. Second, Corollary 3.1 of Vu et al. (2013) does not include the $d^{1/2}\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}$ term in the numerator of the bound. As discussed in the previous paragraph, it is this term that would typically be expected to attain the minimum in (1.2), especially in high-dimensional contexts. We also provide Theorem 1.4 to generalise the result to asymmetric or non-square matrices.

The constants presented in Theorem 1.2 are sharp, as the following example illustrates. Fix $d \in \{1, \ldots, \lfloor p/2 \rfloor\}$ and let $\Sigma = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$, where $\lambda_1 = \cdots = \lambda_{p-2d} = 5$, $\lambda_{p-2d+1} = \cdots = \lambda_{p-d} = 3$ and $\lambda_{p-d+1} = \cdots = \lambda_p = 1$. Suppose that $\hat{\Sigma}$ is also diagonal, with first $p - 2d$ diagonal entries equal to 5, next d diagonal entries equal to 2, and last d diagonal entries equal to $2 + \epsilon$, for some $\epsilon \in (0, 3)$. If we are interested in the middle block of eigenvectors corresponding to those with corresponding eigenvalue 3 in Σ, then for every orthogonal matrix $\hat{O} \in \mathbb{R}^{d \times d}$,

$$\|\hat{V}\hat{O} - V\|_{\mathrm{F}} = 2^{1/2}\|\sin\Theta(\hat{V}, V)\|_{\mathrm{F}} = (2d)^{1/2} \leq (2d)^{1/2}(1 + \epsilon) = \frac{2^{3/2}d^{1/2}\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}}{\delta},$$

where $\delta := \min(\lambda_{p-2d} - \lambda_{p-2d+1}, \lambda_{p-d} - \lambda_{p-d+1})$. In this example, the column spaces of V and $\hat{V}$ were orthogonal. However, even when these column spaces are close, our bound (1.2) is tight up to a factor of 2, while our bound (1.3) is tight up to a factor of $2^{3/2}$. To see this, suppose that $\Sigma = \mathrm{diag}(3, 1)$ while $\hat{\Sigma} = \hat{V}\mathrm{diag}(3, 1)\hat{V}^\top$, where

$$\hat{V} = \begin{pmatrix} (1 - \epsilon^2)^{1/2} & -\epsilon \\ \epsilon & (1 - \epsilon^2)^{1/2} \end{pmatrix}$$

for some $\epsilon > 0$. If $v = (1, 0)^\top$ and $\hat{v} = \big((1 - \epsilon^2)^{1/2}, -\epsilon\big)^\top$ denote the top eigenvectors of Σ and $\hat{\Sigma}$ respectively, then $\sin\Theta(\hat{v}, v) = \epsilon$, $\|\hat{v} - v\|^2 = 2 - 2(1 - \epsilon^2)^{1/2}$ and $2\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}/(3 - 1) = 2\epsilon$.

It is also worth mentioning that there is another theorem in the Davis and Kahan (1970) paper, the so-called $\sin 2\theta$ theorem, which provides a bound for $\|\sin 2\Theta(\hat{V}, V)\|_{\mathrm{F}}$ assuming only a population eigen-gap condition. In the case $d = 1$, this quantity can be related to the square of the length of the difference between the sample and population eigenvectors $\hat{v}$ and v as follows:

$$\sin^2 2\Theta(\hat{v}, v) = (2\hat{v}^\top v)^2\{1 - (\hat{v}^\top v)^2\} = \frac{1}{4}\|\hat{v} - v\|^2(2 - \|\hat{v} - v\|^2)(4 - \|\hat{v} - v\|^2). \tag{1.4}$$

Equation (1.4) reveals, however, that $\|\sin 2\Theta(\hat{V}, V)\|_{\mathrm{F}}$ is unlikely to be of immediate interest to

statisticians, and in fact, we are not aware of applications of the Davis–Kahan $\sin 2\theta$ theorem in statistics. No general bound for $\|\sin\Theta(\hat{V}, V)\|_{\mathrm{F}}$ or $\|\hat{V}\hat{O} - V\|_{\mathrm{F}}$ can be derived from the Davis–Kahan $\sin 2\theta$ theorem since we would require further information such as $\hat{v}^\top v \geq 1/2^{1/2}$ when $d = 1$, and such information would typically be unavailable. The utility of our bound comes from the fact that it provides direct control of the main quantities of interest to statisticians.

Many if not most applications of this result will only need $s = r$, i.e. $d = 1$. In that case, the statement simplifies a little; for ease of reference, we state it as a corollary:

Corollary 1.3. *Let $\Sigma, \hat{\Sigma} \in \mathbb{R}^{p\times p}$ be symmetric, with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$ respectively. Fix $j \in \{1, \ldots, p\}$, and assume that $\min(\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1}) > 0$, where we define $\lambda_0 := \infty$ and $\lambda_{p+1} := -\infty$. If $v, \hat{v} \in \mathbb{R}^p$ satisfy $\Sigma v = \lambda_j v$ and $\hat{\Sigma}\hat{v} = \hat{\lambda}_j \hat{v}$, then*

$$\sin\Theta(\hat{v}, v) \leq \frac{2\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}}{\min(\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1})}.$$

Moreover, if $\hat{v}^\top v \geq 0$, then

$$\|\hat{v} - v\| \leq \frac{2^{3/2}\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}}{\min(\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1})}.$$

1.3 Applications in statistical contexts

In the introduction, we explained how the fact that our variant of the Davis–Kahan $\sin\theta$ theorem only relies on a population eigen-gap condition can be used to simplify many arguments in the statistical literature. These include the work of Fan, Liao and Mincheva (2013) on large covariance matrix estimation problems, Cai, Ma and Wu (2013) on sparse principal component estimation, and Fan and Han (2013) on estimating the false discovery proportion in large-scale multiple testing with highly correlated test statistics. Although our notation suggests that we have covariance matrix estimation in mind, we emphasise that the real, symmetric matrices in Theorem 1.2 are arbitrary, and could be for example inverse covariance matrices, or graph Laplacians as in the work of von Luxburg (2007) and Rohe, Chatterjee and Yu (2011) on spectral clustering in community detection with network data.

We now give some simple examples to illustrate the improvements afforded by our bound in Theorem 1.2. Consider the spiked covariance model in which $X_1, \ldots, X_n$ are independent random vectors having the $N_p(0, \Sigma)$ distribution, where $\Sigma = (\Sigma_{jk})$ is a diagonal matrix with $\Sigma_{jj} = 1 + \theta$ for some $\theta > 0$ for $1 \leq j \leq d$ and $\Sigma_{jj} = 1$ for $d + 1 \leq j \leq p$. Let $\hat{\Sigma} := n^{-1}\sum_{i=1}^n X_i X_i^\top$ denote the sample covariance matrix, and let V and $\hat{V}$ denote the matrices whose columns are unit-length eigenvectors corresponding to the d largest eigenvalues of Σ and $\hat{\Sigma}$ respectively. Fixing $n = 1000$, $p = 200$, $d = 10$ and $\theta = 1$, we found that our bound (1.2) from Theorem 1.2 was an improvement over that from (1.1) in Theorem 1.1 in every one of 100 independent data sets drawn from this model. In fact, no bound could be obtained from Theorem 1 for 25 realisations because δ defined in that result was zero. The median value of $\|\sin\Theta(\hat{V}, V)\|_{\mathrm{F}}$ was 1.80, while the median values of the right-hand sides of (1.2) in Theorem 1.2 and (1.1) in Theorem 1.1 were 7.30 and 376 respectively. Some insight into the reasons for this marked improvement can be gained by considering an asymptotic regime in which $p/n \to \gamma \in (0, 1)$ as $n \to \infty$ and d and θ are considered fixed. Then, in the notation of Theorem 1.1, $\delta = \max(\lambda_d - \hat{\lambda}_{d+1}, 0) \to \max(\theta - 2\gamma^{1/2} - \gamma, 0)$, almost surely, where the limit follows from Baik and Silverstein (2006, Theorem 1.1). On the other

hand, the denominator of the right-hand side of (1.2) in Theorem 1.2 is θ, which may be much larger than $\max(\theta - 2\gamma^{1/2} - \gamma, 0)$. For the numerator, in this example, it can be shown that

$$E(\|\hat{\Sigma} - \Sigma\|_{\mathrm{F}}^2) = \frac{p(p+2)}{n} + \frac{2d(p+2)}{n}\theta + \frac{d(d+2)}{n}\theta^2 \geq \frac{p^2}{n}.$$

Moreover, by Theorem 1.1(b) of Baik, Ben Arous and Péché (2005) and a uniform integrability argument,

$$E(\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}^2) \leq E\{(\hat{\lambda}_1 - 1)^2\} \to \left\{\theta + \frac{(1+\theta)\gamma}{\theta}\right\}^2.$$

We therefore expect the minimum in the numerator of (1.2) to be attained by the term $d^{1/2}\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}$ in this example.

To illustrate our bound in a high-dimensional context, consider the same data generating mechanism as in our previous example. Given an even integer $k \in \{1, \ldots, p\}$, let $\hat{\Sigma} = \hat{\Sigma}_k$ be the tapering estimator for high-dimensional sparse covariance matrices introduced by Cai, Zhang and Zhou (2010). In other words, $\hat{\Sigma}$ is the Hadamard product of the sample covariance matrix and a weight matrix $W = (w_{ij}) \in \mathbb{R}^{p \times p}$, where $w_{ij} = 1$ if $|i - j| \leq k/2$, $w_{ij} = 2 - \frac{2|i-j|}{k}$ if $k/2 < |i - j| < k$ and $w_{ij} = 0$ otherwise. To compare the bounds provided by Theorems 1.1 and 1.2, we drew 100 data sets from this model for each of the settings $n \in \{1000, 2000\}$, $p \in \{2000, 4000\}$, $d = 10$, $\theta = 1$ and $k = 20$. The bound (1.2) improved on that in (1.1) for every realisation in each setting; the medians of these bounds are presented in Table 1.3.

n	p	RHS1	RHS2	n	p	RHS1	RHS2
1000	2000	12.1	2.65	1000	4000	17.3	2.69
2000	2000	7.20	1.92	2000	4000	10.2	1.90

Table 1.1: Median values of RHS1 and RHS2, the bounds obtained from (1.1) and (1.2) respectively.

1.4 Extension to general real matrices

We now describe how the results of Section 1.2 can be extended to situations where the matrices under study may not be symmetric and may not even be square, and where interest is in controlling the principal angles between corresponding singular vectors.

Theorem 1.4. *Suppose that $A, \hat{A} \in \mathbb{R}^{p \times q}$ have singular values $\sigma_1 \geq \cdots \geq \sigma_{\min(p,q)} \geq 0$ and $\hat{\sigma}_1 \geq \cdots \geq \hat{\sigma}_{\min(p,q)}$ respectively. Fix $1 \leq r \leq s \leq \mathrm{rank}(A)$ and assume that $\min(\sigma_{r-1} - \sigma_r, \sigma_s - \sigma_{s+1}) > 0$, where we define $\sigma_0 := \infty$ and $\sigma_{\mathrm{rank}(A)+1} := -\infty$. Let $d := s - r + 1$, and let $V = (v_r, v_{r+1}, \ldots, v_s) \in \mathbb{R}^{q \times d}$ and $\hat{V} = (\hat{v}_r, \hat{v}_{r+1}, \ldots, \hat{v}_s) \in \mathbb{R}^{q \times d}$ have orthonormal columns satisfying $A^\top A v_j = \sigma_j^2 v_j$ and $\hat{A}^\top \hat{A} \hat{v}_j = \hat{\sigma}_j^2 \hat{v}_j$ for $j = r, r+1, \ldots, s$. Then*

$$\|\sin\Theta(\hat{V}, V)\|_{\mathrm{F}} \leq \frac{4\min(d^{1/2}\|\hat{A} - A\|_{\mathrm{op}}, 2^{1/2}\|\hat{A} - A\|_{\mathrm{F}})}{\min(\sigma_{r-1} - \sigma_r, \sigma_s - \sigma_{s+1})}. \tag{1.5}$$

Moreover, there exists an orthogonal matrix $\hat{O} \in \mathbb{R}^{d \times d}$ such that

$$\|\hat{V}\hat{O} - V\|_{\mathrm{F}} \leq \frac{2^{5/2}\min(d^{1/2}\|\hat{A} - A\|_{\mathrm{op}}, 2^{1/2}\|\hat{A} - A\|_{\mathrm{F}})}{\min(\sigma_{r-1} - \sigma_r, \sigma_s - \sigma_{s+1})}.$$

Theorem 1.4 gives bounds on the proximity of the right singular vectors of A and $\hat{A}$. Identical bounds also hold if V and $\hat{V}$ are replaced with the matrices of left singular vectors U and $\hat{U}$, where $U = (u_r, u_{r+1}, \ldots, u_s) \in \mathbb{R}^{p\times d}$ and $\hat{U} = (\hat{u}_r, \hat{u}_{r+1}, \ldots, \hat{u}_s) \in \mathbb{R}^{p\times d}$ have orthonormal columns satisfying $AA^\top u_j = \sigma_j^2 u_j$ and $\hat{A}\hat{A}^\top \hat{u}_j = \hat{\sigma}_j^2 \hat{u}_j$ for $j = r, r+1, \ldots, s$.

As mentioned in the introduction, Theorem 1.4 can be viewed as a variant of the generalised $\sin\theta$ theorem of Wedin (1972). Similar to the situation for symmetric matrices, there are many places in the statistical literature where Wedin's result has been used, but where we argue that Theorem 1.4 above would be a more natural result to which to appeal. Examples include the papers of Van Huffel and Vandewalle (1989) on the accuracy of least squares techniques, Anandkumar et al. (2014) on tensor decompositions for learning latent variable models, Shabalin and Nobel (2013) on recovering a low-rank matrix from a noisy version and Sun and Zhang (2012) on matrix completion.

1.5 Appendix

We first state an elementary lemma that will be useful in several places.

Lemma 1.5. *Let $A \in \mathbb{R}^{m\times n}$, and let $U \in \mathbb{R}^{m\times p}$ and $W \in \mathbb{R}^{n\times q}$ both have orthonormal rows. Then $\|U^\top AW\|_{\mathrm{F}} = \|A\|_{\mathrm{F}}$. If instead $U \in \mathbb{R}^{m\times p}$ and $W \in \mathbb{R}^{n\times q}$ both have orthonormal columns, then $\|U^\top AW\|_{\mathrm{F}} \leq \|A\|_{\mathrm{F}}$.*

Proof. For the first claim,

$$\|U^\top AW\|_{\mathrm{F}}^2 = \mathrm{tr}(U^\top AWW^\top A^\top U) = \mathrm{tr}(AA^\top UU^\top) = \mathrm{tr}(AA^\top) = \|A\|_{\mathrm{F}}^2.$$

For the second part, find a matrix $U_1 \in \mathbb{R}^{m\times(m-p)}$ such that $\begin{pmatrix} U & U_1 \end{pmatrix}$ is orthogonal, and a matrix $W_1 \in \mathbb{R}^{n\times(n-q)}$ such that $\begin{pmatrix} W & W_1 \end{pmatrix}$ is orthogonal. Then

$$\|A\|_{\mathrm{F}} = \left\| \begin{pmatrix} U^\top \\ U_1^\top \end{pmatrix} A \begin{pmatrix} W & W_1 \end{pmatrix} \right\|_{\mathrm{F}} \geq \left\| \begin{pmatrix} U^\top \\ U_1^\top \end{pmatrix} AW \right\|_{\mathrm{F}} \geq \|U^\top AW\|_{\mathrm{F}}.$$

as desired. □

Proof of Theorem 1.2. Define matrices $\Lambda := \mathrm{diag}(\lambda_r, \lambda_{r+1}, \ldots, \lambda_s)$ and $\hat{\Lambda} := \mathrm{diag}(\hat{\lambda}_r, \ldots, \hat{\lambda}_s)$. Then

$$0 = \hat{\Sigma}\hat{V} - \hat{V}\hat{\Lambda} = \Sigma\hat{V} - \hat{V}\Lambda + (\hat{\Sigma} - \Sigma)\hat{V} - \hat{V}(\hat{\Lambda} - \Lambda).$$

Hence

$$\begin{aligned} \|\hat{V}\Lambda - \Sigma\hat{V}\|_{\mathrm{F}} &\leq \|(\hat{\Sigma} - \Sigma)\hat{V}\|_{\mathrm{F}} + \|\hat{V}(\hat{\Lambda} - \Lambda)\|_{\mathrm{F}} \\ &\leq d^{1/2}\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}} + \|\hat{\Lambda} - \Lambda\|_{\mathrm{F}} \leq 2d^{1/2}\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}, \end{aligned} \tag{1.6}$$

where we have used Lemma 1.5 in the second inequality and Weyl's inequality (e.g. Stewart and Sun, 1990, Corollary IV.4.9) for the final bound. Alternatively, we can argue that

$$\begin{aligned} \|\hat{V}\Lambda - \Sigma\hat{V}\|_{\mathrm{F}} &\leq \|(\hat{\Sigma} - \Sigma)\hat{V}\|_{\mathrm{F}} + \|\hat{V}(\hat{\Lambda} - \Lambda)\|_{\mathrm{F}} \\ &\leq \|\hat{\Sigma} - \Sigma\|_{\mathrm{F}} + \|\hat{\Lambda} - \Lambda\|_{\mathrm{F}} \leq 2\|\hat{\Sigma} - \Sigma\|_{\mathrm{F}}, \end{aligned} \tag{1.7}$$

where the second inequality follows from two applications of Lemma 1.5, and the final inequality follows from the Wielandt–Hoffman theorem (e.g. Wilkinson, 1965, pp. 104–108).

Let $\Lambda_1 = \mathrm{diag}(\lambda_1, \ldots, \lambda_{r-1}, \lambda_{s+1}, \ldots, \lambda_p)$, and let V_1 be a $p \times (p-d)$ matrix such that $P = \begin{pmatrix} V & V_1 \end{pmatrix}$ is orthogonal and such that

$$P^\top \Sigma P = \begin{pmatrix} \Lambda & 0 \\ 0 & \Lambda_1 \end{pmatrix}.$$

Then

$$\begin{aligned} \|\hat{V}\Lambda - \Sigma\hat{V}\|_{\mathrm{F}} &= \|VV^\top\hat{V}\Lambda + V_1V_1^\top\hat{V}\Lambda - V\Lambda V^\top\hat{V} - V_1\Lambda_1V_1^\top\hat{V}\|_{\mathrm{F}} \\ &\geq \|V_1V_1^\top\hat{V}\Lambda - V_1\Lambda_1V_1^\top\hat{V}\|_{\mathrm{F}} \geq \|V_1^\top\hat{V}\Lambda - \Lambda_1V_1^\top\hat{V}\|_{\mathrm{F}}, \end{aligned} \tag{1.8}$$

where the first inequality follows because $V^\top V_1 = 0$, and the second from another application of Lemma 1.5. For real matrices A and B, we write $A \otimes B$ for their Kronecker product (e.g. Stewart and Sun, 1990, p. 30) and $\mathrm{vec}(A)$ for the vectorisation of A, i.e. the vector formed by stacking its columns. We recall the standard identity $\mathrm{vec}(ABC) = (C^\top \otimes A)\mathrm{vec}(B)$, which holds whenever the dimensions of the matrices are such that the matrix multiplication is well-defined. We also write I_m for the m-dimensional identity matrix. Then

$$\begin{aligned} \|V_1^\top\hat{V}\Lambda - \Lambda_1V_1^\top\hat{V}\|_{\mathrm{F}} &= \|(\Lambda \otimes I_{p-d} - I_d \otimes \Lambda_1)\mathrm{vec}(V_1^\top\hat{V})\| \\ &\geq \min(\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1})\|\mathrm{vec}(V_1^\top\hat{V})\| \\ &= \min(\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1})\|\sin\Theta(\hat{V}, V)\|_{\mathrm{F}}, \end{aligned} \tag{1.9}$$

where the final step follows from

$$\begin{aligned} \|\mathrm{vec}(V_1^\top\hat{V})\|^2 = \mathrm{tr}(\hat{V}^\top V_1V_1^\top\hat{V}) = \mathrm{tr}\big((I_p - VV^\top)\hat{V}\hat{V}^\top\big) &= d - \|\hat{V}^\top V\|_{\mathrm{F}}^2 \\ &= \|\sin\Theta(\hat{V}, V)\|_{\mathrm{F}}^2. \end{aligned}$$

Now (1.2) follows from (1.9), (1.8), (1.7) and (1.6).

For the second conclusion, by a singular value decomposition, we can find orthogonal matrices $\hat{O}_1, \hat{O}_2 \in \mathbb{R}^{d\times d}$ such that

$$\hat{O}_1^\top\hat{V}^\top V\hat{O}_2 = \mathrm{diag}(\cos\theta_1, \ldots, \cos\theta_d),$$

where $\theta_1, \ldots, \theta_d$ are the principal angles between the column spaces of V and $\hat{V}$. Setting $\hat{O} = \hat{O}_1\hat{O}_2^\top$, we have

$$\begin{aligned} \|\hat{V}\hat{O} - V\|_{\mathrm{F}}^2 &= \mathrm{tr}\big((\hat{V}\hat{O} - V)^\top(\hat{V}\hat{O} - V)\big) = 2d - 2\mathrm{tr}(\hat{O}_2\hat{O}_1^\top\hat{V}^\top V) \\ &= 2d - 2\sum_{j=1}^d \cos\theta_j \leq 2d - 2\sum_{j=1}^d \cos^2\theta_j = 2\|\sin\Theta(\hat{V}, V)\|_{\mathrm{F}}^2. \end{aligned}$$

The result now follows from our first conclusion. □

Proof of Theorem 1.4. Let $u_1, \ldots, u_{\min(p,q)} \in \mathbb{R}^p$ and $v_1, \ldots, v_{\min(p,q)} \in \mathbb{R}^q$ be the two sets of orthonormal vectors that are left and right singular vectors of A corresponding to the singular values $\sigma_1, \ldots, \sigma_{\min(p,q)}$. In other words, we have $Av_j = \sigma_j u_j$ and $A^\top u_j = \sigma_j v_j$. Define similarly

$\hat{u}_1, \dots, \hat{u}_{\min(p,q)}$ and $\hat{v}_1, \dots, \hat{v}_{\min(p,q)}$ to be left and right singular vectors of $\hat{A}$ corresponding to singular values $\hat{\sigma}_1 \geq \cdots \geq \hat{\sigma}_{\min(p,q)} \geq 0$. We symmetrise A and $\hat{A}$ by defining

$$B = \begin{pmatrix} 0 & A^\top \\ A & 0 \end{pmatrix} \quad \text{and} \quad \hat{B} = \begin{pmatrix} 0 & \hat{A}^\top \\ \hat{A} & 0 \end{pmatrix}.$$

We can check that B is symmetric with eigenvalues $\pm\sigma_1, \dots, \pm\sigma_{\min(p,q)}, 0, \dots, 0$ and if we define $w_j := 2^{-1/2}(v_j^\top, u_j^\top)^\top$, then $w_1, \dots, w_{\min(p,q)}$ are orthonormal and satisfy $Bw_j = \sigma_j w_j$. Similarly, if we define $\hat{w}_j := 2^{-1/2}(\hat{v}_j^\top, \hat{u}_j^\top)^\top$, then $\hat{w}_1, \dots, \hat{w}_{\min(p,q)}$ are orthonormal and $\hat{B}\hat{w}_j = \hat{\sigma}_j \hat{w}_j$.

Let $W = (w_r, \dots, w_s) \in \mathbb{R}^{(q+p)\times d}$ and $\hat{W} = (\hat{w}_r, \dots, \hat{w}_s) \in \mathbb{R}^{(q+p)\times d}$. Then by Theorem 1.2, we have

$$\|\sin\Theta(\hat{W}, W)\|_{\mathrm{F}} \leq \frac{2\min(d^{1/2}\|\hat{B} - B\|_{\mathrm{op}}, \|\hat{B} - B\|_{\mathrm{F}})}{\min(\sigma_{r-1} - \sigma_r, \sigma_s - \sigma_{s+1})}. \tag{1.10}$$

On the other hand, we can bound

$$\begin{aligned} \|\sin\Theta(\hat{W}, W)\|_{\mathrm{F}}^2 &= d - \|\hat{W}^\top W\|_{\mathrm{F}}^2 = d - \frac{1}{4}\|\hat{V}^\top V + \hat{U}^\top U\|_{\mathrm{F}}^2 \\ &\geq \left(\frac{d}{2} - \frac{\|\hat{V}^\top V\|_{\mathrm{F}}^2}{2}\right) + \left(\frac{d}{2} - \frac{\|\hat{U}^\top U\|_{\mathrm{F}}^2}{2}\right) \\ &= \frac{1}{2}\big(\|\sin\Theta(\hat{V}, V)\|_{\mathrm{F}}^2 + \|\sin\Theta(\hat{U}, U)\|_{\mathrm{F}}^2\big). \end{aligned} \tag{1.11}$$

The bound in (1.5) then follows from (1.10), (1.11) and the facts that $\|\hat{B} - B\|_{\mathrm{op}} = \|\hat{A} - A\|_{\mathrm{op}}$ and $\|\hat{B} - B\|_{\mathrm{F}} = 2^{1/2}\|\hat{A} - A\|_{\mathrm{F}}$. Finally, the bound for $\|\hat{V}\hat{O} - V\|_{\mathrm{F}}$ follows immediately from (1.5) as in the proof of Theorem 1.2. □

Chapter 2

Statistical and computational trade-offs in estimation of sparse principal components

2.1 Introduction

Principal Component Analysis (PCA), which involves projecting a sample of multivariate data onto the space spanned by the leading eigenvectors of the sample covariance matrix, is one of the oldest and most widely-used dimension reduction devices in Statistics. It has proved to be particularly effective when the dimension of the data is relatively small by comparison with the sample size. However, the work of Johnstone and Lu (2009) and Paul (2007) shows that PCA breaks down in the high-dimensional settings that are frequently encountered in many diverse modern application areas. For instance, consider the spiked covariance model where $X_1, \ldots, X_n$ are independent $N_p(0, \Sigma)$ random vectors, with $\Sigma = I_p + \theta v_1 v_1^\top$ for some $\theta > 0$ and an arbitrary unit vector $v_1 \in \mathbb{R}^p$. In this case, v_1 is the leading eigenvector (principal component) of Σ, and the classical PCA estimate would be $\hat{v}_1$, a unit-length leading eigenvector of the sample covariance matrix $\hat{\Sigma} := n^{-1} \sum_{i=1}^n X_i X_i^\top$. In the high-dimensional setting where $p = p_n$ is such that $p/n \to c \in (0, 1)$, Paul (2007) showed that

$$|\hat{v}_1^\top v_1| \overset{a.s.}{\to} \begin{cases} \sqrt{\frac{1-c/\theta^2}{1+c/\theta}} & \text{if } \theta > \sqrt{c} \\ 0 & \text{if } \theta \leq \sqrt{c}. \end{cases}$$

In other words, $\hat{v}_1$ is inconsistent as an estimator of v_1 in this asymptotic regime. This phenomenon is related to the so-called 'BBP' transition in random matrix theory (Baik, Ben Arous and Péché, 2005).

Sparse Principal Component Analysis was designed to remedy this inconsistency and to give additional interpretability to the projected data. In the simplest case, it is assumed that the leading eigenvector v_1 of the population covariance matrix Σ belongs to the k-sparse unit Euclidean sphere in $\mathbb{R}^p$, given by

$$\mathbb{S}^{p-1}(k) := \left\{ u = (u_1, \ldots, u_p)^\top \in \mathbb{R}^p : \sum_{j=1}^p \mathbb{1}_{\{u_j \neq 0\}} \leq k, \ \|u\|_2 = 1 \right\}. \tag{2.1}$$

A remarkable number of recent papers have proposed estimators of v_1 in this setting, including Jolliffe, Trendafilov and Uddin (2003), Zou, Hastie and Tibshirani (2006), d'Aspremont et al. (2007), Johnstone and Lu (2009), Witten, Tibshirani and Hastie (2009), Journée et al. (2010), Birnbaum et al. (2013), Cai, Ma and Wu (2013), Ma (2013), Shen, Shen and Marron (2013) and Vu and Lei (2013).

Sparse PCA methods have gained high popularity in many diverse applied fields where high-dimensional datasets are routinely handled. These include computer vision for online visual tracking (Wang, Lu and Yang, 2013) and pattern recognition (Naikal, Yang and Sastry, 2011), signal processing for image compression (Majumdar, 2009) and Electrocardiography feature extraction (Johnstone and Lu, 2009), and biomedical research for gene expression analysis (Zou, Hastie and Tibshirani, 2006; Chun and Sündüz, 2009; Parkhomenko, Tritchler and Beyene, 2009; Chan and Hall, 2010), RNA-seq classification (Tan, Petersen and Witten, 2014) and metabolomics studies (Genevera and Maletić-Savatić, 2011). In these applications, Sparse PCA is employed to identify a small number of interpretable directions that represent the data succinctly, typically as the first stage of a more involved procedure such as classification, clustering or regression.

The success of the ultimate inferential methods in the types of application described above depends critically on how well the particular Sparse PCA technique involved identifies the relevant meaningful directions in the underlying population. It therefore becomes important to understand the ways in which our ability to estimate these directions from data depends on the characteristics of the problem, including the sample size, dimensionality, sparsity level and signal-to-noise ratio. Such results form a key component of any theoretical analysis of an inference problem in which Sparse PCA is employed as a first step.

In terms of the theoretical properties of existing methods for Sparse PCA, Ma (2013) was able to show that his estimator attains the minimax rate of convergence over a certain Gaussian class of distributions, provided that k is treated as a fixed constant. Both Cai, Ma and Wu (2013) and Vu and Lei (2013) also study minimax properties, but treat k as a parameter of the problem that may vary with the sample size n. In particular, for a certain class $\mathcal{P}_p(n,k)$ of subgaussian distributions and in a particular asymptotic regime, Vu and Lei (2013) show[1] that

$$\inf_{\hat{v}} \sup_{P \in \mathcal{P}_p(n,k)} \mathbb{E}_P\{1 - (v_1^\top \hat{v})^2\} \asymp \frac{k \log p}{n},$$

where the infimum is taken over all estimators $\hat{v}$; see also Birnbaum et al. (2013). Moreover, they show that the minimax rate is attained by a leading k-sparse eigenvector of $\hat{\Sigma}$, given by

$$\hat{v}^k_{\max} \in \operatorname*{argmax}_{u \in \mathbb{S}^{p-1}(k)} u^\top \hat{\Sigma} u. \tag{2.2}$$

The papers cited above would appear to settle the question of sparse principal component estimation (at least in a subgaussian setting) from the perspective of statistical theory. However, there remains an unsettling feature, namely that neither the estimator of Cai, Ma and Wu (2013), nor that of Vu and Lei (2013), is computable in polynomial time[2]. For instance, computing the estimator (2.2) is an NP-hard problem, and the naive algorithm that searches through all $\binom{p}{k}$ of the $k \times k$ principal submatrices of $\hat{\Sigma}$ quickly becomes infeasible for even moderately large p and k.

Given that Sparse PCA is typically applied to massive high-dimensional datasets, it is crucial

[1] Here and below, $a_n \asymp b_n$ means $0 < \liminf_{n\to\infty} |a_n/b_n| \leq \limsup_{n\to\infty} |a_n/b_n| < \infty$.

[2] To keep the thesis as self-contained as possible, a brief introduction to this topic is provided in Section 2.8.

to understand the rates that can be achieved using only computationally efficient procedures. Specifically, in this chapter, we address the question of whether it is possible to find an estimator of v_1 that is computable in (randomised) polynomial time, and that attains the minimax optimal rate of convergence when the sparsity of v_1 is allowed to vary with the sample size. Some progress in a related direction was made by Berthet and Rigollet (2013a,b), who considered the problem of testing the null hypothesis $H_0 : \Sigma = I_p$ against the alternative $H_1 : v^\top \Sigma v \geq 1 + \theta$ for some $v \in \mathbb{S}^{p-1}(k)$ and $\theta > 0$. Of interest here is the minimal level $\theta = \theta_{n,p,k}$ that ensures small asymptotic testing error. Under a hypothesis on the computational intractability of a certain well-known problem from theoretical computer science (the 'Planted Clique' detection problem), Berthet and Rigollet showed that for certain classes of distributions, there is a gap between the minimal θ-level permitting successful detection with a randomised polynomial time test, and the corresponding θ-level when arbitrary tests are allowed.

The particular classes of distributions considered in Berthet and Rigollet (2013a,b) were highly tailored to the testing problem, and do not provide sufficient structure to study principal component estimation. The thesis of the current chapter, however, is that from the point of view of both theory and applications, it is the estimation of sparse principal components, rather than testing for the existence of a distinguished direction, that is the more natural and fundamental (as well as more challenging) problem. Indeed, we observe subtle phase transition phenomena that are absent from the hypothesis testing problem; see Section 2.4.4 for further details. It is worth noting that different results for statistical and computational trade-offs for estimation and testing were also observed in the context of k-SAT formulas in Feldman, Perkins and Vempala (2015) and Berthet (2015).

Our first contribution, in Section 2.2, is to introduce a new Restricted Covariance Concentration (RCC) condition that underpins the classes of distributions $\mathcal{P}_p(n, k, \theta)$ over which we perform the statistical and computational analyses (see (2.4) for a precise definition). The RCC condition is satisfied by subgaussian distributions, and moreover has the advantage of being more robust to certain mixture contaminations that turn out to be of key importance in the statistical analysis under the computational constraint. We show that subject to mild restrictions on the parameter values,

$$\inf_{\hat{v}} \sup_{P \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_P L(\hat{v}, v_1) \asymp \sqrt{\frac{k \log p}{n\theta^2}},$$

where $L(u, v) := \{1 - (u^\top v)^2\}^{1/2}$, and where no restrictions are placed on the class of estimators $\hat{v}$. By contrast, in Section 2.3, we show that a variant $\hat{v}^{\mathrm{SDP}}$ of the semidefinite relaxation estimator of d'Aspremont et al. (2007) and Bach, Ahipaşaoğlu and d'Aspremont (2010), which is computable in polynomial time, satisfies

$$\sup_{P \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_P L(\hat{v}^{\mathrm{SDP}}, v_1) \leq (16\sqrt{2} + 2)\sqrt{\frac{k^2 \log p}{n\theta^2}}.$$

Our main result, in Section 2.4, is that, under a much weaker Planted Clique hypothesis than that in Berthet and Rigollet (2013a,b), for any $\alpha \in (0, 1)$, there exists a moderate effective sample size asymptotic regime in which every sequence $(\hat{v}^{(n)})$ of randomised polynomial time estimators satisfies

$$\sqrt{\frac{n\theta^2}{k^{1+\alpha} \log p}} \sup_{P \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_P L(\hat{v}^{(n)}, v_1) \to \infty.$$

This result shows that there is a fundamental trade-off between statistical and computational

efficiency in the estimation of sparse principal components, and that there is in general no consistent sequence of randomised polynomial time estimators in this regime. Interestingly, in a high effective sample size regime, where even randomised polynomial time estimators can be consistent, we are able to show in Theorem 2.7 that under additional distributional assumptions, a modified (but still polynomial time) version of $\hat{v}^{\mathrm{SDP}}$ attains the minimax optimal rate. Thus, the trade-off disappears for a sufficiently high effective sample size, at least over a subset of the parameter space.

Statistical and computational trade-offs have also recently been studied in the context of convex relaxation algorithms (Chandrasekaran and Jordan, 2013), submatrix signal detection (Ma and Wu, 2015; Chen and Xu, 2016), sparse linear regression (Zhang, Wainwright and Jordan, 2014), community detection (Hajek, Wu and Xu, 2015) and Sparse Canonical Correlation Analysis (Gao, Ma and Zhou, 2014). Given the importance of computationally feasible algorithms with good statistical performance in today's era of Big Data, it seems clear that understanding the extent of this phenomenon in different settings will represent a key challenge for theoreticians in the coming years.

Proofs of our main results are given in Section 2.6, while several ancillary results are deferred to Section 2.7. We end this section by introducing some notation used throughout this chapter. We write $\mathbb{S}^{p-1}$ for the unit sphere in $\mathbb{R}^p$. For a vector $u = (u_1, \ldots, u_M)^\top \in \mathbb{R}^M$, a matrix $A = (A_{ij}) \in \mathbb{R}^{M \times N}$ and for $q \in [1, \infty)$, we write $\|u\|_q := \big(\sum_{i=1}^M |u_i|^q\big)^{1/q}$ and $\|A\|_q := \big(\sum_{i=1}^M \sum_{j=1}^N |A_{ij}|^q\big)^{1/q}$ for their (entrywise) ℓ_q-norms. We also write $\|u\|_0 := \sum_{i=1}^M \mathbb{1}_{\{u_i \neq 0\}}$, $\mathrm{supp}(u) := \{i : u_i \neq 0\}$, $\|A\|_0 := \sum_{i=1}^M \sum_{j=1}^N \mathbb{1}_{\{A_{ij} \neq 0\}}$ and $\mathrm{supp}(A) := \{(i,j) : A_{ij} \neq 0\}$. For $S \subseteq \{1, \ldots, M\}$ and $T \subseteq \{1, \ldots, N\}$, we write $u_S := (u_i : i \in S)^\top$ and write $M_{S,T}$ for the $|S| \times |T|$ submatrix of M obtained by extracting the rows and columns with indices in S and T respectively. For positive sequences (a_n) and (b_n), we write $a_n \ll b_n$ to mean $a_n / b_n \to 0$.

2.2 Restricted Covariance Concentration and minimax rate of estimation

Let $p \geq 2$ and let $\mathcal{P}$ denote the class of probability distributions P on $\mathbb{R}^p$ with $\int_{\mathbb{R}^p} x \, dP(x) = 0$ and such that the entries of $\Sigma(P) := \int_{\mathbb{R}^p} xx^\top \, dP(x)$ are finite. For $P \in \mathcal{P}$, write $\lambda_1(P), \ldots, \lambda_p(P)$ for the eigenvalues of $\Sigma(P)$, arranged in decreasing order. When $\lambda_1(P) - \lambda_2(P) > 0$, the first principal component $v_1(P)$, i.e. a unit-length eigenvector of Σ corresponding to the eigenvalue $\lambda_1(P)$, is well-defined up to sign. In some places below, and where it is clear from the context, we suppress the dependence of these quantities on P, or write the eigenvalues and eigenvectors as $\lambda_1(\Sigma), \ldots, \lambda_p(\Sigma)$ and $v_1(\Sigma), \ldots, v_p(\Sigma)$ respectively. Let $X_1, \ldots, X_n$ be independent and identically distributed random vectors with distribution P, and form the $n \times p$ matrix $\mathbf{X} := (X_1, \ldots, X_n)^\top$. An *estimator* of v_1 is a measurable function from $\mathbb{R}^{n \times p}$ to $\mathbb{R}^p$, and we write $\mathcal{V}_{n,p}$ for the class of all such estimators.

Given unit vectors $u, v \in \mathbb{R}^p$, let $\Theta(u, v) := \cos^{-1}(|u^\top v|)$ denote the acute angle between u and v, and define the loss function

$$L(u, v) := \sin \Theta(u, v) = \{1 - (u^\top v)^2\}^{1/2} = \frac{1}{\sqrt{2}} \|uu^\top - vv^\top\|_2.$$

Note that $L(\cdot, \cdot)$ is invariant to sign changes of either of its arguments. The *directional variance* of P along a unit vector $u \in \mathbb{R}^p$ is defined to be $V(u) := \mathbb{E}\{(u^\top X_1)^2\} = u^\top \Sigma u$. Its empirical version

is $\hat{V}(u) := n^{-1}\sum_{i=1}^n (u^\top X_i)^2 = u^\top \hat{\Sigma} u$, where $\hat{\Sigma} := n^{-1}\sum_{i=1}^n X_i X_i^\top$ denotes the sample covariance matrix.

Recall the definition of the k-sparse unit ball $\mathbb{S}^{p-1}(k)$ from (2.1). Given $\ell \in \{1,\dots,p\}$ and $C \in (0,\infty)$, we say P satisfies a *Restricted Covariance Concentration* (RCC) condition with parameters p, n, ℓ and C, and write $P \in \mathrm{RCC}_p(n,\ell,C)$, if

$$\mathbb{P}\Big\{ \sup_{u\in\mathbb{S}^{p-1}(\ell)} |\hat{V}(u) - V(u)| \geq C\max\Big(\sqrt{\frac{\ell\log(p/\delta)}{n}}, \frac{\ell\log(p/\delta)}{n}\Big)\Big\} \leq \delta \tag{2.3}$$

for all $\delta > 0$. It is also convenient to define

$$\mathrm{RCC}_p(\ell,C) := \bigcap_{n=1}^{\infty} \mathrm{RCC}_p(n,\ell,C) \quad \text{and} \quad \mathrm{RCC}_p(C) := \bigcap_{\ell=1}^{p} \mathrm{RCC}_p(\ell,C).$$

The RCC conditions amount to uniform Bernstein-type concentration properties of the directional variance around its expectation along all sparse directions. This condition turns out to be particularly convenient in the study of convergence rates in Sparse PCA, and moreover, as we show in Proposition 2.1 below, subgaussian distributions satisfy an RCC condition for all sample sizes n and all sparsity levels ℓ. Recall that a mean-zero distribution Q on $\mathbb{R}^p$ is *subgaussian* with parameter[3] $\sigma^2 \in (0,\infty)$, written $Q \in \mathrm{subgaussian}_p(\sigma^2)$, if whenever $Y \sim Q$, we have $\mathbb{E}(e^{u^\top Y}) \leq e^{\sigma^2\|u\|^2/2}$ for all $u \in \mathbb{R}^p$.

Proposition 2.1. *(i) For every $\sigma > 0$, we have* $\mathrm{subgaussian}_p(\sigma^2) \subseteq \mathrm{RCC}_p\big(16\sigma^2(1+9/\log p)\big)$.
(ii) In the special case where $P = N_p(0,\Sigma)$, we have $P \in \mathrm{RCC}_p\big(8\lambda_1(P)(1+\frac{9}{\log p})\big)$.

Our convergence rate results for sparse principal component estimation will be proved over the following classes of distributions. For $\theta > 0$, let

$$\mathcal{P}_p(n,k,\theta) := \Big\{ P \in \mathrm{RCC}_p(n,2,1) \cap \mathrm{RCC}_p(n,2k,1) : v_1(P) \in \mathbb{S}^{p-1}(k), \lambda_1(P) - \lambda_2(P) \geq \theta \Big\}. \tag{2.4}$$

Observe that RCC classes have the scaling property that if the distribution of a random vector Y belongs to $\mathrm{RCC}_p(n,\ell,C)$ and if $r > 0$, then the distribution of rY belongs to $\mathrm{RCC}_p(n,\ell,r^2C)$. It is therefore convenient to fix $C = 1$ in both RCC classes in (2.4), so that θ becomes a measure of the signal-to-noise level.

For a symmetric $A \in \mathbb{R}^{p\times p}$, define $\hat{v}^k_{\max}(A) := \mathrm{sargmax}_{u\in\mathbb{S}^{p-1}(k)} u^\top A u$ to be the k-sparse maximum eigenvector of A, where sargmax denotes the smallest element of the argmax in the lexicographic ordering. (This choice ensures that $\hat{v}^k_{\max}(A)$ is a measurable function of A.) Theorem 2.2 below gives a finite-sample minimax upper bound for estimating $v_1(P)$ over $\mathcal{P}_p(n,k,\theta)$. For similar bounds over Gaussian or subgaussian classes, see Cai, Ma and Wu (2013) and Vu and Lei (2013), who consider the more general problem of principal subspace estimation. As well as working with a larger class of distributions, our different proof techniques facilitate an explicit constant.

Theorem 2.2. *For $2k\log p \leq n$, the k-sparse empirical maximum eigenvector, $\hat{v}^k_{\max}(\hat{\Sigma})$, satisfies*

$$\sup_{P\in\mathcal{P}_p(n,k,\theta)} \mathbb{E}_P L\big(\hat{v}^k_{\max}(\hat{\Sigma}), v_1(P)\big) \leq 2\sqrt{2}\Big(1+\frac{1}{\log p}\Big)\sqrt{\frac{k\log p}{n\theta^2}} \leq 7\sqrt{\frac{k\log p}{n\theta^2}}.$$

[3]Note that some authors say that distributions satisfying this condition are subgaussian with parameter σ, rather than σ^2.

A matching minimax lower bound of the same order in all parameters k, p, n and θ is given below. The proof techniques are adapted from Vu and Lei (2013).

Theorem 2.3. *Suppose that* $7 \leq k \leq p^{1/2}$ *and* $0 < \theta \leq \frac{1}{16(1+\frac{9}{\log p})}$. *Then*

$$\inf_{\hat{v} \in \mathcal{V}_{n,p}} \sup_{P \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_P L\big(\hat{v}, v_1(P)\big) \geq \min\left\{ \frac{1}{1660}\sqrt{\frac{k \log p}{n\theta^2}}, \frac{5}{18\sqrt{3}} \right\}.$$

We remark that the conditions in the statement of Theorem 2.3 can be strengthened or weakened, with a corresponding weakening or strengthening of the constants in the bound. For instance, a bound of the same order in k, p, n and θ could be obtained assuming only that $k \leq p^{1-\delta}$ for some $\delta > 0$. The upper bound on θ is also not particularly restrictive. For example, if $P = N_p(0, \sigma^2 I_p + \theta e_1 e_1^\top)$, where e_1 is the first standard basis vector in $\mathbb{R}^p$, then it can be shown that the condition $P \in \mathcal{P}_p(n, k, \theta)$ requires that $\theta \leq 1 - \sigma^2$.

2.3 Computationally efficient estimation

As was mentioned in the introduction, the trouble with the estimator $\hat{v}^k_{\max}(\hat{\Sigma})$ of Section 2.2, as well as the estimator of Cai, Ma and Wu (2013), is that there are no known polynomial time algorithms for their computation. In this section, we therefore study the (polynomial time) semidefinite relaxation estimator $\hat{v}^{\mathrm{SDP}}$ defined by the Algorithm 2.1 below. This estimator is a variant of one proposed by d'Aspremont et al. (2007), whose support recovery properties were studied for a particular class of Gaussian distributions and a known sparsity level by Amini and Wainwright (2009).

To motivate the main step (Step 2) of Algorithm 2.1, it is convenient to let $\mathcal{M}$ denote the class of $p \times p$ non-negative definite real, symmetric matrices, and let $\mathcal{M}_1 := \{M \in \mathcal{M} : \mathrm{tr}(M) = 1\}$. Let $\mathcal{M}_{1,1}(k^2) := \{M \in \mathcal{M}_1 : \mathrm{rank}(M) = 1, \|M\|_0 = k^2\}$ and observe that

$$\max_{u \in \mathbb{S}^{p-1}(k)} u^\top \hat{\Sigma} u = \max_{u \in \mathbb{S}^{p-1}(k)} \mathrm{tr}(\hat{\Sigma} u u^\top) = \max_{M \in \mathcal{M}_{1,1}(k^2)} \mathrm{tr}(\hat{\Sigma} M).$$

In the final expression, the rank and sparsity constraints are non-convex. We therefore adopt the standard semidefinite relaxation approach of dropping the rank constraint and replacing the sparsity (ℓ_0) constraint with an ℓ_1 penalty to obtain the convex optimisation problem

$$\max_{M \in \mathcal{M}_1} \big\{ \mathrm{tr}(\hat{\Sigma} M) - \lambda \|M\|_1 \big\} \tag{2.5}$$

Algorithm 2.1: Pseudo-code for computing the semidefinite relaxation estimator $\hat{v}^{\mathrm{SDP}}$

Input: $\mathbf{X} = (X_1, \ldots, X_n)^\top \in \mathbb{R}^{n \times p}$, $\lambda > 0, \epsilon > 0$
begin
  Step 1: Set $\hat{\Sigma} \leftarrow n^{-1}\mathbf{X}^\top \mathbf{X}$.
  Step 2: For $f(M) := \mathrm{tr}(\hat{\Sigma} M) - \lambda \|M\|_1$, let $\hat{M}^\epsilon$ be an ϵ-maximiser of f in $\mathcal{M}_1$. In other words, $\hat{M}^\epsilon$ satisfies $f(\hat{M}^\epsilon) \geq \max_{M \in \mathcal{M}_1} f(M) - \epsilon$.
  Step 3: Let $\hat{v}^{\mathrm{SDP}} := \hat{v}^{\mathrm{SDP}}_{\lambda,\epsilon} \in \mathrm{argmax}_{u : \|u\|_2 = 1} u^\top \hat{M}^\epsilon u$.
end
Output: $\hat{v}^{\mathrm{SDP}}$

We now discuss the complexity of computing $\hat{v}^{\mathrm{SDP}}$ in detail. One possible way of implementing Step 2 is to use a generic interior-point method. However, as shown in Nesterov (2005), Nemirovski (2004) and Bach, Ahipaşaoğlu and d'Aspremont (2010), certain first-order algorithms (i.e. methods requiring $O(1/\epsilon)$ steps to find a feasible point achieving an ϵ-approximation of the optimal objective function value) can significantly outperform such generic interior-point solvers. The key idea in both Nesterov (2005) and Nemirovski (2004) is that the optimisation problem in Step 2 can be rewritten in a saddlepoint formulation:

$$\max_{M \in \mathcal{M}_1} \operatorname{tr}(\hat{\Sigma} M) - \lambda \|M\|_1 = \max_{M \in \mathcal{M}_1} \min_{U \in \mathcal{U}} \operatorname{tr}\big((\hat{\Sigma} + U)M\big),$$

where $\mathcal{U} := \{U \in \mathbb{R}^{p \times p} : U^\top = U, \|U\|_\infty \leq \lambda\}$. The fact that $\operatorname{tr}\big((\hat{\Sigma} + U)M\big)$ is linear in both M and U makes the problem amenable to proximal methods. In Algorithm 2.2 below, we state a possible implementation of Step 2 of Algorithm 2.1, derived from the 'basic implementation' in Nemirovski (2004). In the algorithm, the $\|\cdot\|_2$-norm projection $\Pi_{\mathcal{U}}(A)$ of a symmetric matrix $A = (A_{ij}) \in \mathbb{R}^{p \times p}$ onto $\mathcal{U}$ is given by $\big(\Pi_{\mathcal{U}}(A)\big)_{ij} := \operatorname{sign}(A_{ij}) \min(|A_{ij}|, \lambda)$. For the projection $\Pi_{\mathcal{M}_1}(A)$, first decompose $A =: PDP^\top$ for some orthogonal P and diagonal $D = \operatorname{diag}(d)$, where $d = (d_1, \ldots, d_p)^\top \in \mathbb{R}^p$. Now let $\Pi_{\mathcal{W}}(d)$ be the projection image of d on the unit $(p-1)$-simplex $\mathcal{W} := \{(w_1, \ldots, w_p) : w_j \geq 0, \sum_{j=1}^p w_j = 1\}$. Finally, transform back to obtain $\Pi_{\mathcal{M}_1}(A) := P \operatorname{diag}\big(\Pi_{\mathcal{W}}(d)\big) P^\top$. The fact that Algorithm 2.2 outputs an ϵ-maximiser of the optimisation problem in Step 2 of Algorithm 2.1 is a consequence of Nemirovski (2004, Theorem 3.2), which implies in our particular case that after N iterations,

$$\max_{M \in \mathcal{M}_1} \min_{U \in \mathcal{U}} \operatorname{tr}\big((\hat{\Sigma} + U)M\big) - \min_{U \in \mathcal{U}} \operatorname{tr}\big((\hat{\Sigma} + U)\hat{M}^\epsilon\big) \leq \frac{\lambda^2 p^2 + 1}{\sqrt{2}N}.$$

In Algorithm 2.1, Step 1 takes $O(np^2)$ floating point operations; Step 3 takes $O(p^3)$ operations in the worst case, though other methods such as the Lanczos method (Lanczos, 1950; Golub and Van Loan, 1996) require only $O(p^2)$ operations under certain conditions. Our particular implementation (Algorithm 2.2) for Step 2 requires $O(\frac{\lambda^2 p^2 + 1}{\epsilon})$ iterations in the worst case, though this number may often be considerably reduced by terminating the **for** loop if the primal-dual gap

$$\lambda_1(\hat{U}_t + \hat{\Sigma}) - \{\operatorname{tr}(\hat{M}_t \hat{\Sigma}) - \lambda \|\hat{M}_t\|_1\}$$

falls below ϵ, where $\hat{U}_t := t^{-1} \sum_{s=1}^t U'_s$ and $\hat{M}_t := t^{-1} \sum_{s=1}^t M'_s$. The most costly step within the **for** loop is the eigendecomposition used to compute the projection $\Pi_{\mathcal{M}_1}$, which takes $O(p^3)$ operations. Taking $\lambda := 4\sqrt{\frac{\log p}{n}}$ and $\epsilon := \frac{\log p}{4n}$ as in Theorem 2.5 below, we find an overall complexity for the algorithm of $O\big(\max(p^5, \frac{np^3}{\log p})\big)$ operations in the worst case.

We now turn to the theoretical properties of the estimator $\hat{v}^{\mathrm{SDP}}$ computed using Algorithm 2.1. Lemma 2.4 below is stated in a general, deterministic fashion, but will be used in Theorem 2.5 below to bound the loss incurred by the estimator on the event that the sample and population covariance matrices are close in ℓ_∞-norm. See also Vu et al. (2013, Theorem 3.1) for a closely related result in the context of a projection matrix estimation problem. Recall that $\mathcal{M}$ denotes the class of $p \times p$ non-negative definite real, symmetric matrices.

Lemma 2.4. *Let $\Sigma \in \mathcal{M}$ be such that $\theta := \lambda_1(\Sigma) - \lambda_2(\Sigma) > 0$. Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $\hat{\Sigma} := n^{-1}\mathbf{X}^\top \mathbf{X}$. For arbitrary $\lambda > 0$ and $\epsilon > 0$, if $\|\hat{\Sigma} - \Sigma\|_\infty \leq \lambda$, then the semidefinite relaxation estimator $\hat{v}^{\mathrm{SDP}}$*

Algorithm 2.2: A possible implementation of Step 2 of Algorithm 2.1

Input: $\hat{\Sigma} \in \mathcal{M}$, $\lambda > 0$, $\epsilon > 0$.
begin
  Set $M_0 \leftarrow I_p/p$, $U_0 \leftarrow 0 \in \mathbb{R}^{p\times p}$ and $N \leftarrow \left\lceil \frac{\lambda^2 p^2 + 1}{\sqrt{2}\epsilon} \right\rceil$.
  for $t \leftarrow 1$ **to** N **do**
    $U'_t \leftarrow \Pi_{\mathcal{U}}\big(U_{t-1} - \frac{1}{\sqrt{2}} M_{t-1}\big)$, $M'_t \leftarrow \Pi_{\mathcal{M}_1}\big(M_{t-1} + \frac{1}{\sqrt{2}}\hat{\Sigma} + \frac{1}{\sqrt{2}} U_{t-1}\big)$.
    $U_t \leftarrow \Pi_{\mathcal{U}}\big(U_{t-1} - \frac{1}{\sqrt{2}} M'_t\big)$, $M_t \leftarrow \Pi_{\mathcal{M}_1}\big(M_{t-1} + \frac{1}{\sqrt{2}}\hat{\Sigma} + \frac{1}{\sqrt{2}} U'_t\big)$.
  end
  Set $\hat{M}^\epsilon \leftarrow \frac{1}{N}\sum_{t=1}^N M'_t$.
end
Output: $\hat{M}^\epsilon$

in Algorithm 2.1 with inputs $\mathbf{X}, \lambda, \epsilon$ *satisfies*

$$L\big(\hat{v}^{\mathrm{SDP}}, v_1(\Sigma)\big) \le \frac{4\sqrt{2}\lambda k}{\theta} + 2\sqrt{\frac{\epsilon}{\theta}}.$$

Theorem 2.5 below describes the statistical properties of the estimator $\hat{v}^{\mathrm{SDP}}$ over the classes $\mathcal{P}_p(n, k, \theta)$. It reveals in particular that we incur a loss of statistical efficiency of a factor of $\sqrt{k}$ compared with the minimax upper bound in Theorem 2.2 in Section 2.2 above. As well as applying Lemma 2.4 on the event $\{\|\hat{\Sigma} - \Sigma\|_\infty \le \lambda\}$, the proof relies on Lemma 2.12 in Section 2.7, which relates the event $\{\|\hat{\Sigma} - \Sigma\|_\infty > \lambda\}$ to the $\mathrm{RCC}_p(n, 2, 1)$ condition. Indeed, this explains why we incorporated this condition into the definition of the $\mathcal{P}_p(n, k, \theta)$ classes.

Theorem 2.5. *For an arbitrary* $P \in \mathcal{P}_p(n, k, \theta)$ *and* $X_1, \ldots, X_n \overset{iid}{\sim} P$, *we write* $\hat{v}^{\mathrm{SDP}}(\mathbf{X})$ *for the output of Algorithm 2.1 with input* $\mathbf{X} := (X_1, \ldots, X_n)^\top$, $\lambda := 4\sqrt{\frac{\log p}{n}}$ *and* $\epsilon := \frac{\log p}{4n}$. *If* $4\log p \le n \le k^2 p^2 \theta^{-2} \log p$ *and* $\theta \in (0, k]$, *then*

$$\sup_{P \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_P L\big(\hat{v}^{\mathrm{SDP}}(\mathbf{X}), v_1(P)\big) \le \min\left\{(16\sqrt{2} + 2)\sqrt{\frac{k^2 \log p}{n\theta^2}}, 1\right\}. \tag{2.6}$$

We remark that $\hat{v}^{\mathrm{SDP}}$ has the attractive property of being fully adaptive in the sense that it can be computed without knowledge of the sparsity level k. On the other hand, $\hat{v}^{\mathrm{SDP}}$ is not necessarily k-sparse. If a specific sparsity level is desired in a particular application, Algorithm 2.1 can be modified to obtain a (non-adaptive) k-sparse estimator having similar estimation risk. Specifically, we can find

$$\hat{v}_0^{\mathrm{SDP}} \in \operatorname*{argmin}_{u \in \mathbb{S}^{p-1}(k)} L(\hat{v}^{\mathrm{SDP}}, u).$$

Since $L(\hat{v}^{\mathrm{SDP}}, u)^2 = 1 - \big(u^\top \hat{v}^{\mathrm{SDP}}\big)^2$, we can compute $\hat{v}_0^{\mathrm{SDP}}$ by setting all but the top k coordinates of $\hat{v}^{\mathrm{SDP}}$ in absolute value to zero and renormalising the vector. In particular, $\hat{v}_0^{\mathrm{SDP}}$ is computable in polynomial time. We deduce that under the same conditions as in Theorem 2.5, for any $P \in \mathcal{P}_p(n, k, \theta)$,

$$\begin{aligned}
\mathbb{E}L\big(\hat{v}_0^{\mathrm{SDP}}, v_1\big) &\le \mathbb{E}\big[\big\{L\big(\hat{v}_0^{\mathrm{SDP}}, \hat{v}^{\mathrm{SDP}}\big) + L\big(\hat{v}^{\mathrm{SDP}}, v_1\big)\big\}\mathbb{1}_{\{\|\hat{\Sigma}-\Sigma\|_\infty \le \lambda\}}\big] + \mathbb{P}\big(\|\hat{\Sigma} - \Sigma\|_\infty > \lambda\big) \\
&\le 2\mathbb{E}\big\{L\big(\hat{v}_0^{\mathrm{SDP}}, v_1\big)\mathbb{1}_{\{\|\hat{\Sigma}-\Sigma\|_\infty \le \lambda\}}\big\} + \mathbb{P}\big(\|\hat{\Sigma} - \Sigma\|_\infty > \lambda\big) \le (32\sqrt{2} + 3)\sqrt{\frac{k^2 \log p}{n\theta^2}},
\end{aligned}$$

where the final inequality follows from the proof of Theorem 2.5.

2.4 Computational lower bounds

Theorems 2.5 and 2.2 reveal a gap between the provable performance of our semidefinite relaxation estimator $\hat{v}^{\mathrm{SDP}}$ and the minimax optimal rate. It is natural to ask whether there exists a computationally efficient algorithm that achieves the statistically optimal rate of convergence. In fact, as we will see in Theorem 2.6 below, the effective sample size region over which $\hat{v}^{\mathrm{SDP}}$ is consistent is essentially tight among the class of all *randomised polynomial time algorithms*[4]. Indeed, any randomised polynomial time algorithm with a faster rate of convergence could otherwise be adapted to solve instances of the Planted Clique problem that are believed to be hard; see Section 2.4.1 below for formal definitions and discussion. In this sense, the extra factor of $\sqrt{k}$ is an intrinsic price in statistical efficiency that we have to pay for computational efficiency, and the estimator $\hat{v}^{\mathrm{SDP}}$ studied in Section 2.3 has essentially the best possible rate of convergence among computable estimators.

2.4.1 The Planted Clique problem

A *graph* $G := (V(G), E(G))$ is an ordered pair in which $V(G)$ is a countable set, and $E(G)$ is a subset of $\big\{\{x, y\} : x, y \in V(G), x \neq y\big\}$. For $x, y \in V(G)$, we say x and y are *adjacent*, and write $x \sim y$, if $\{x, y\} \in E(G)$. A *clique* C is a subset of $V(G)$ such that $\{x, y\} \in E(G)$ for all distinct $x, y \in C$. The problem of finding a clique of maximum size in a given graph G is known to be NP-*complete* (Karp, 1972). It is therefore natural to consider randomly generated input graphs with a clique 'planted' in, where the signal is much less confounded by the noise. Such problems were first suggested by Jerrum (1992) and Kučera (1995) as a potentially easier variant of the classical Clique problem.

Let $\mathbb{G}_m$ denote the collection of all graphs with m vertices. Define $\mathcal{G}_m$ to be the distribution on $\mathbb{G}_m$ associated with the standard Erdős–Rényi random graph. In other words, under $\mathcal{G}_m$, each pair of vertices is adjacent independently with probability $1/2$. For any $\kappa \in \{1, \ldots, m\}$, let $\mathcal{G}_{m,\kappa}$ be a distribution on $\mathbb{G}_m$ constructed by first picking κ distinct vertices uniformly at random and connecting all edges (the 'planted clique'), then joining each remaining pair of distinct vertices by an edge independently with probability $1/2$. The Planted Clique problem has input graphs randomly sampled from the distribution $\mathcal{G}_{m,\kappa}$. Due to the random nature of the problem, the goal of the Planted Clique problem is to find (possibly randomised) algorithms that can locate a maximum clique K_m with high probability.

It is well known that, for a standard Erdős–Rényi graph, $\frac{|K_m|}{2\log_2 m} \xrightarrow{\text{a.s.}} 1$ (e.g. Grimmett and McDiarmid, 1975). In fact, if $\kappa = \kappa_m$ is such that $\liminf_{m\to\infty} \frac{\kappa}{2\log_2 m} > 1$, it can be shown that the planted clique is asymptotically almost surely also the unique maximum clique in the input graph. As observed in Kučera (1995), there exists $C > 0$ such that, if $\kappa > C\sqrt{m \log m}$, then asymptotically almost surely, vertices in the planted clique have larger degrees than all other vertices, in which case they can be located in $O(m^2)$ operations. Alon, Krivelevich and Sudakov (1998) improved the above result by exhibiting a spectral method that, given any $c > 0$, identifies planted cliques of size $\kappa \geq c\sqrt{m}$ asymptotically almost surely.

[4]In this section, terms from computational complexity theory defined Section 2.8 are written in italics at their first occurrence.

Although several other polynomial time algorithms have subsequently been discovered for the $\kappa \geq c\sqrt{m}$ case (e.g. Feige and Krauthgamer, 2000; Feige and Ron, 2010; Ames and Vavasis, 2011), there is no known randomised polynomial time algorithm that can detect below this threshold. Jerrum (1992) hinted at the hardness of this problem by showing that a specific Markov chain approach fails to work when $\kappa = O(m^{1/2-\delta})$ for some $\delta > 0$. Feige and Krauthgamer (2003) showed that Lovàcz–Schrijiver semidefinite programming relaxation methods also fail in this regime. Feldman et al. (2013) recently presented further evidence of the hardness of this problem by showing that a broad class of algorithms, which they refer to as 'statistical algorithms', cannot solve the Planted Clique problem with $\kappa = O(m^{1/2-\delta})$ in randomised polynomial time, for any $\delta > 0$. It is now widely accepted in theoretical computer science that the Planted Clique problem is hard, in the sense that the following assumption holds with $\tau = 0$:

(A1)(τ) For any sequence $\kappa = \kappa_m$ such that $\kappa \leq m^\beta$ for some $0 < \beta < 1/2 - \tau$, there is no randomised polynomial time algorithm that can correctly identify the planted clique with probability tending to 1 as $m \to \infty$.

We state the assumption in terms of a general parameter $\tau \in [0, 1/2)$, because it will turn out below that even if only (A1)(τ) holds for some $\tau \in (0, 1/6)$, there are still regimes of (n, p, k, θ) in which no randomised polynomial time algorithm can attain the minimax optimal rate.

Researchers have used the hardness of the planted clique problem as an assumption to prove various impossibility results in other problems. Examples include cryptographic applications (Juels and Peinado, 2000; Applebaum, Barak and Wigderson, 2010), testing k-wise independence (Alon et al., 2007) and approximating Nash equilibria (Hazan and Krauthgamer, 2011). Recent works by Berthet and Rigollet (2013a,b) and Ma and Wu (2015) used a stronger hypothesis on the hardness of detecting the presence of a planted clique to establish computational lower bounds in sparse principal component detection and sparse submatrix detection problems respectively. Our Assumption (A1)(0) assumes only the computational intractability of identifying the entire planted clique, so in particular, is implied by Hypothesis $\mathrm{A_{PC}}$ of Berthet and Rigollet (2013b) and Hypothesis 1 of Ma and Wu (2015).

2.4.2 Main theorem

In this section, we use a reduction argument to show that, under Assumption (A1)(τ), it is impossible to achieve the statistically optimal rate of sparse principal component estimation using randomised polynomial time algorithms. For $\rho \in \mathbb{N}$, and for $x \in \mathbb{R}$, we let $[x]_\rho$ denote x in its binary representation, rounded to ρ significant figures. Let $[\mathbb{R}]_\rho := \{[x]_\rho : x \in \mathbb{R}\}$. We say $(\hat{v}^{(n)})$ is a *sequence of randomised polynomial time estimators* of $v_1 \in \mathbb{R}^{p_n}$ if $\hat{v}^{(n)}$ is a measurable function from $\mathbb{R}^{n \times p_n}$ to $\mathbb{R}^{p_n}$ and if, for every $\rho \in \mathbb{N}$, there exists a randomised polynomial time algorithm M_{pr} such that for any $\mathbf{x} \in ([\mathbb{R}]_\rho)^{n \times p_n}$ we have $[\hat{v}^{(n)}(\mathbf{x})]_\rho = [M_{\mathrm{pr}}(\mathbf{x})]_\rho$. The sequence of semidefinite programming estimators $(\hat{v}^{\mathrm{SDP}})$ defined in Section 2.3 is an example of a sequence of randomised polynomial time estimators of $v_1(P)$.

Theorem 2.6. *Fix $\tau \in [0, 1/6)$, assume (A1)(τ), and let $\alpha \in (0, \frac{1-6\tau}{1-2\tau})$. For any $n \in \mathbb{N}$, let $(p, k, \theta) = (p_n, k_n, \theta_n)$ be parameters indexed by n such that $k = O(p^{1/2-\tau-\delta})$ for some $\delta \in (0, 1/2-\tau)$, $n = o(p \log p)$ and $\theta \leq k^2/(1000p)$. Suppose further that*

$$\frac{k^{1+\alpha} \log p}{n\theta^2} \to 0$$

as $n \to \infty$. Let $\mathbf{X}$ be an $n \times p$ matrix with independent rows, each having distribution P. Then every sequence $(\hat{v}^{(n)})$ of randomised polynomial time estimators of $v_1(P)$ satisfies

$$\sqrt{\frac{n\theta^2}{k^{1+\alpha}\log p}} \sup_{P \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_P L\big(\hat{v}^{(n)}(\mathbf{X}), v_1(P)\big) \to \infty$$

as $n \to \infty$.

We note that the choices of parameters in the theorem imply that

$$\liminf_{n\to\infty} \frac{k^2 \log p}{n\theta^2} \geq \liminf_{n\to\infty} \frac{p}{k^2} = \infty. \tag{2.7}$$

As remarked in Section 2.4.1 above, the main interest in this theorem comes from the case $\tau = 0$. Here, our result reveals not only that no randomised polynomial time algorithm can attain the minimax optimal rate, but also that in the effective sample size regime described by (2.7), and provided the other side conditions of Theorem 2.6 hold, there is in general no consistent sequence of randomised polynomial time estimators. This is in contrast to Theorem 2.2, where we saw that consistent estimation with a computationally inefficient procedure is possible in the asymptotic regime (2.7). A further consequence of Theorem 2.6 is that, since any sequence $(p, k, \theta) = (p_n, k_n, \theta_n)$ satisfying the conditions of Theorem 2.6 also satisfies the conditions of Theorem 2.5 for large n, the conclusion of Theorem 2.5 cannot be improved in terms of the exponent of k (at least, not uniformly over the parameter range given there). As mentioned in the introduction, for a sufficiently large effective sample size, where even randomised polynomial time estimators can be consistent, the statistical and computational trade-off revealed by Theorems 2.2 and 2.6 may disappear. See Section 2.4.4 below for further details, and Gao, Ma and Zhou (2014) for recent extensions of these results to different classes of distributions.

Even though Assumption (A1)(0) is widely believed, we also present results under the weaker family of conditions (A1)(τ) for $\tau \in (0, 1/6)$ to show that a statistical and computational trade-off still remains for certain parameter regimes even in these settings. The reason for assuming $\tau < 1/6$ is to guarantee that there is a regime of parameters (n, p, k, θ) satisfying the conditions of the theorem. Indeed, if $\tau \in [0, 1/6)$ and $\alpha \in (0, \frac{1-6\tau}{1-2\tau})$, we can set $p = n$, $k = n^{1/2-\tau-\delta}$ for some $\delta \in \big(0, \frac{1}{2} - \tau - \frac{1}{3-\alpha}\big)$, $\theta = k^2/(1000n)$, and in that case,

$$\frac{k^{1+\alpha}\log p}{n\theta^2} = \frac{10^6 n \log n}{k^{3-\alpha}} \to 0,$$

as required.

2.4.3 Sketch of the proof of the main theorem

The proof of Theorem 2.6 relies on a randomised polynomial time reduction from the Planted Clique problem to the sparse principal component estimation problem. The reduction is adapted from the 'bottom-left transformation' of Berthet and Rigollet (2013b), and requires a rather different and delicate analysis.

In greater detail, suppose for a contradiction that we were given a randomised polynomial time algorithm $\hat{v}$ for the Sparse PCA problem with a rate $\sup_{P \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_P L(\hat{v}, v_1) \leq \sqrt{\frac{k^{1+\alpha}\log p}{n\theta^2}}$ for some $\alpha < 1$. Set $m \approx p \log p$ and $\kappa \approx k \log p$, so we are in the regime where (A1)(τ) holds. Given

any graph $G \sim \mathcal{G}_{m,\kappa}$ with planted clique $K \subseteq V(G)$, we draw $n+p$ vertices $u_1, \ldots, u_n, w_1, \ldots, w_p$ uniformly at random without replacement from $V(G)$. On average there are about $\kappa/\log\kappa$ clique vertices in $\{w_1, \ldots, w_p\}$, and our initial aim is to identify a large fraction of these vertices. To do this, we form an $n \times p$ matrix $\mathbf{A} := (\mathbb{1}_{u_i \sim w_j})_{i,j}$, which is an off-diagonal block of the adjacency matrix of G. We then replace each 0 in $\mathbf{A}$ with -1 and flip the signs of each row independently with probability $1/2$ to obtain a new matrix $\mathbf{X}$. Each component of the ith row of $\mathbf{X}$ has a marginal Rademacher distribution, but if u_i is a clique vertex, then the components $\{j : w_j \in K\}$ are perfectly correlated. Writing $\boldsymbol{\gamma}' := (\mathbb{1}_{\{w_j \in K\}})_{j=1,\ldots,p}$, the leading eigenvector of $\mathbb{E}\{\mathbf{X}^\top\mathbf{X}/n|\boldsymbol{\gamma}'\}$ is proportional to $\boldsymbol{\gamma}'$, which suggests that a spectral method might be able to find $\{w_1, \ldots, w_p\} \cap K$ with high probability. Unfortunately, the joint distribution of the rows of $\mathbf{X}$ is difficult to deal with directly, but since n and p are small relative to m, we can approximate $\boldsymbol{\gamma}'$ by a random vector $\boldsymbol{\gamma}$ having independent $\mathrm{Bern}(\kappa/m)$ components. We can then approximate $\mathbf{X}$ by a matrix $\mathbf{Y}$, whose rows are independent conditional on $\boldsymbol{\gamma}$ and have the same marginal distribution conditional on $\boldsymbol{\gamma} = g$ as the rows of $\mathbf{X}$ conditional on $\boldsymbol{\gamma}' = g$.

It turns out that the distribution of an appropriately scaled version of an arbitrary row of $\mathbf{Y}$, conditional on $\boldsymbol{\gamma} = g$, belongs to $\mathcal{P}_p(n, k, \theta)$ for g belonging to a set of high probability. We could therefore apply our hypothetical randomised polynomial time Sparse PCA algorithm to the scaled version of the matrix $\mathbf{Y}$ to find a good estimate of $\boldsymbol{\gamma}$, and since $\boldsymbol{\gamma}$ is close to $\boldsymbol{\gamma}'$, this accomplishes our initial goal. With high probability, the remaining vertices in the planted clique are those having high connectivity to the identified clique vertices in $\{w_1, \ldots, w_p\}$, which contradicts the hypothesis (A1)(τ).

2.4.4 Computationally efficient optimal estimation on a subparameter space in the high effective sample size regime

Theorems 2.2, 2.3, 2.5 and 2.6 enable us to summarise, in Table 2.1 below, our knowledge of the best possible rate of estimation in different asymptotic regimes, both for arbitrary statistical procedures and for those that are computable in randomised polynomial time. (For ease of exposition, we omit here the additional, relatively mild, side constraints required for the above theorems to hold.) The fact that Theorem 2.6 is primarily concerned with the setting in which $\frac{k^2 \log p}{n\theta^2} \to \infty$ raises the question of whether computationally efficient procedures could attain a faster rate of convergence in the high effective sample size regime where $n \gg \frac{k^2 \log p}{\theta^2}$.

The purpose of this section is to extend the ideas of Amini and Wainwright (2009) to show that, indeed, a variant of the estimator $\hat{v}^{\mathrm{SDP}}$ introduced in Section 2.3 attains the minimax optimal rate of convergence in this asymptotic regime, at least over a subclass of the distributions in $\mathcal{P}_p(n, k, \theta)$. Ma (2013) and Yuan and Zhang (2013) show similar results for an iterative thresholding algorithm for other subclasses of $\mathcal{P}_p(n, k, \theta)$ under an extra upper bound condition on $\lambda_2(P)/\lambda_1(P)$; see also Wang, Lu and Liu (2014) and Deshpande and Montanari (2014).

Let $\mathcal{T}$ denote the set of non-negative definite matrices $\Sigma \in \mathbb{R}^{p \times p}$ of the form

$$\Sigma = \theta v_1 v_1^\top + \begin{pmatrix} I_k & 0 \\ 0 & \Gamma_{p-k} \end{pmatrix},$$

where $v_1 \in \mathbb{R}^p$ is a unit vector such that $S := \mathrm{supp}(v_1)$ has cardinality k and where $\Gamma_{p-k} \in \mathbb{R}^{(p-k)\times(p-k)}$ is non-negative definite and satisfies $\lambda_1(\Gamma_{p-k}) \leq 1$. (Here, and in the proof of

Table 2.1: Rates of convergence of in different asymptotic regimes

	$n \ll \frac{k \log p}{\theta^2}$	$\frac{k \log p}{\theta^2} \ll n \ll \frac{k^2 \log p}{\theta^2}$	$n \gg \frac{k^2 \log p}{\theta^2}$
all estimators	$\asymp 1$	$\asymp \sqrt{\frac{k \log p}{n\theta^2}}$	$\asymp \sqrt{\frac{k \log p}{n\theta^2}}$
poly-time estimators	$\asymp 1$	$\asymp 1$	$\lesssim \sqrt{\frac{k^2 \log p}{n\theta^2}}$

Theorem 2.7 below, the block matrix notation refers to the (S, S), (S, S^c), (S^c, S) and (S^c, S^c) blocks.) We now define a subclass of distributions

$$\tilde{\mathcal{P}}_p(n, k, \theta) := \Big\{ P \in \mathcal{P}_p(n, k, \theta) \,:\, \Sigma(P) \in \mathcal{T}, \min_{j \in S} |v_{1,j}| \geq 16 \sqrt{\frac{k \log p}{n\theta^2}} \Big\}.$$

We remark that $\tilde{\mathcal{P}}_p(n, k, \theta)$ is non-empty only if $\sqrt{\frac{k^2 \log p}{n\theta^2}} \leq \frac{1}{16}$, since

$$1 = \|v_{1,S}\|_2 \geq k^{1/2} \min_{j \in S} |v_{1,j}| \geq 16 \sqrt{\frac{k^2 \log p}{n\theta^2}}.$$

This is one reason that the theorem below only holds in the high effective sample size regime. Our variant of $\hat{v}^{\mathrm{SDP}}$ is described in Algorithm 2.3 below. We remark that $\hat{v}^{\mathrm{MSDP}}$, like $\hat{v}^{\mathrm{SDP}}$, is computable in polynomial time.

Algorithm 2.3: Pseudo-code for computing the modified semidefinite relaxation estimator $\hat{v}^{\mathrm{MSDP}}$

Input: $\mathbf{X} = (X_1, \ldots, X_n)^\top \in \mathbb{R}^{n \times p}$, $\lambda > 0$, $\epsilon > 0$, $\tau > 0$.
begin
Step 1: Set $\hat{\Sigma} \leftarrow n^{-1} \mathbf{X}^\top \mathbf{X}$.
Step 2: For $f(M) := \mathrm{tr}(\hat{\Sigma} M) - \lambda \|M\|_1$, let $\hat{M}^\epsilon$ be an ϵ-maximiser of f in $\mathcal{M}_1$. **Step 3:** Let $\hat{S} \leftarrow \big\{ j \in \{1, \ldots, p\} : \hat{M}^\epsilon_{jj} \geq \tau \big\}$ and $\hat{v}^{\mathrm{MSDP}} \in \mathbb{R}^p$ by $\hat{v}^{\mathrm{MSDP}}_{\hat{S}^c} \leftarrow 0$ and $\hat{v}^{\mathrm{MSDP}}_{\hat{S}} \in \mathrm{argmax}_{u \in \mathbb{R}^{|\hat{S}|}} \, u^\top \hat{\Sigma}_{\hat{S}\hat{S}} u$.
end
Output: $\hat{v}^{\mathrm{MSDP}}$

Theorem 2.7. *Assume that* $X_1, \ldots, X_n \overset{\mathrm{iid}}{\sim} P$ *for some* $P \in \tilde{\mathcal{P}}_p(n, k, \theta)$.

(a) Let $\lambda := 4\sqrt{\frac{\log p}{n}}$. *The function* f *in Step 2 of Algorithm 2.3 has a maximiser* $\hat{M} \in \mathcal{M}_{1,1}(k^2)$ *satisfying* $\mathrm{sgn}(\hat{M}) = \mathrm{sgn}(v_1 v_1^\top)$.

(b) Assume that $\log p \leq n$, $\theta^2 \leq B k^{1/2}$ *for some* $B \geq 1$ *and* $p \geq \theta (n/k)^{1/2}$. *We write* $\hat{v}^{\mathrm{MSDP}}$ *for the output of Algorithm 2.3 with input parameters* $\mathbf{X} := (X_1, \ldots, X_n)^\top \in \mathbb{R}^{n \times p}$, $\lambda := 4\sqrt{\frac{\log p}{n}}$, $\epsilon := (\frac{\log p}{Bn})^{5/2}$ *and* $\tau := (\frac{\log p}{Bn})^2$. *Then*

$$\sup_{P \in \tilde{\mathcal{P}}_p(n,k,\theta)} \mathbb{E}_P \big\{ L(\hat{v}^{\mathrm{MSDP}}, v_1) \big\} \leq 6 \sqrt{\frac{k \log p}{n\theta^2}}.$$

Theorem 2.7 generalises Theorem 2 of Amini and Wainwright (2009) in two ways: first, we relax a gaussianity assumption to an RCC condition; second, the leading eigenvector of the population covariance matrix is not required to have non-zero entries equal to $\pm k^{-1/2}$.

2.5 Numerical experiments

In this section we present the results of numerical experiments to illustrate the results of Theorems 2.5, 2.6 and 2.7. We generate $v_1 \in \mathbb{R}^p$ by setting $v_{1,j} := k^{-1/2}$ for $j = 1, \ldots, k$, and $v_{1,j} := 0$ for $j = k+1, \ldots, p$. We then draw $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N_p(0, \Sigma)$, where $\Sigma := I_p + \theta v_1 v_1^\top$ and $\theta = 1$. We apply Algorithm 2.1 to the data matrix $\mathbf{X} := (X_1, \ldots, X_n)^\top$ and report the average loss of the estimator $\hat{v}^{\text{SDP}}$ over $N_{\text{rep}} := 100$ repetitions. For $p \in \{50, 100, 150, 200\}$ and $k = \lfloor p^{1/2} \rfloor$, we repeat the experiment for several choices of n to explore the three parameter regimes described in Table 2.1. Since the boundaries of these regimes are $n \asymp \frac{k \log p}{\theta^2}$ and $n \asymp \frac{k^2 \log p}{\theta^2}$, we plot the average loss of the experiments against effective samples sizes

$$\nu_{\text{lin}} := \frac{n\theta^2}{k \log p}, \quad \text{and} \quad \nu_{\text{quad}} := \frac{n\theta^2}{k^2 \log p}.$$

The results are shown in Figure 2.1. The top left panel of Figure 2.1 shows a sharp phase transition for the average loss, as predicted by Theorems 2.5 and 2.6. The right panels of Figure 2.1 suggest that in the high effective sample size regime, $\hat{v}^{\text{SDP}}$ converges at rate $\sqrt{\frac{k \log p}{n\theta^2}}$ in this setting. This is the same rate as was proved for the modified semidefinite relaxation estimator $\hat{v}^{\text{MSDP}}$ in Theorem 2.7.

It is worth noting that it is relatively time-consuming to carry out the simulations for the settings in the right-hand tails of the plots in Figure 2.1. These extreme settings were chosen, however, to illustrate that the linear scaling is the correct one in this tail. For example, when $\nu_{\text{quad}} = 200$ and $p = 200$, we require $n = 207694$, and the pre-processing of the data matrix to obtain the sample covariance matrix is the time-limiting step. In general, in our experience, the semi-definite programming algorithm is certainly not as fast as simpler methods such as diagonal thresholding, but is not prohibitively slow.

2.6 Appendix: Proofs of the main results

Proof of Proposition 2.1. (i) If $X_1, \ldots, X_n \overset{iid}{\sim} P$ for $P \in \text{subgaussian}_p(\sigma^2)$, then, for any $u \in \mathbb{S}^{p-1}(\ell)$ and $t \geq 0$, we have a tail probability bound $\mathbb{P}(u^\top X_1 \geq t) \leq e^{-t^2/\sigma^2} \mathbb{E}(e^{tu^\top X_1/\sigma^2}) \leq e^{-t^2/(2\sigma^2)}$. Similarly, $\mathbb{P}(-u^\top X_1 \geq t) \leq e^{-t^2/(2\sigma^2)}$. Write $\mu_u := \mathbb{E}\{(u^\top X_1)^2\}$; since

$$1 + \frac{1}{2}\mu_u t^2 + o(t^2) = \mathbb{E}(e^{tu^\top X_1}) \leq e^{t^2\sigma^2/2} = 1 + \frac{1}{2}\sigma^2 t^2 + o(t^2),$$

as $t \to 0$, we deduce that $\mu_u \leq \sigma^2$. Now, for any integer $m \geq 2$,

$$\begin{aligned}
\mathbb{E}\big(\big|(u^\top X_1)^2 - \mu_u\big|^m\big) &\leq \int_0^\infty \mathbb{P}\Big\{(u^\top X_1)^2 - \mu_u \geq t^{1/m}\Big\}\, dt + \mu_u^m \\
&\leq 2\int_0^\infty e^{-\frac{t^{1/m} + \mu_u}{2\sigma^2}}\, dt + \mu_u^m = m!(2\sigma^2)^m \left\{2e^{-\mu_u/(2\sigma^2)} + \frac{1}{m!}\Big(\frac{\mu_u}{2\sigma^2}\Big)^m\right\} \leq 2m!(2\sigma^2)^m,
\end{aligned}$$

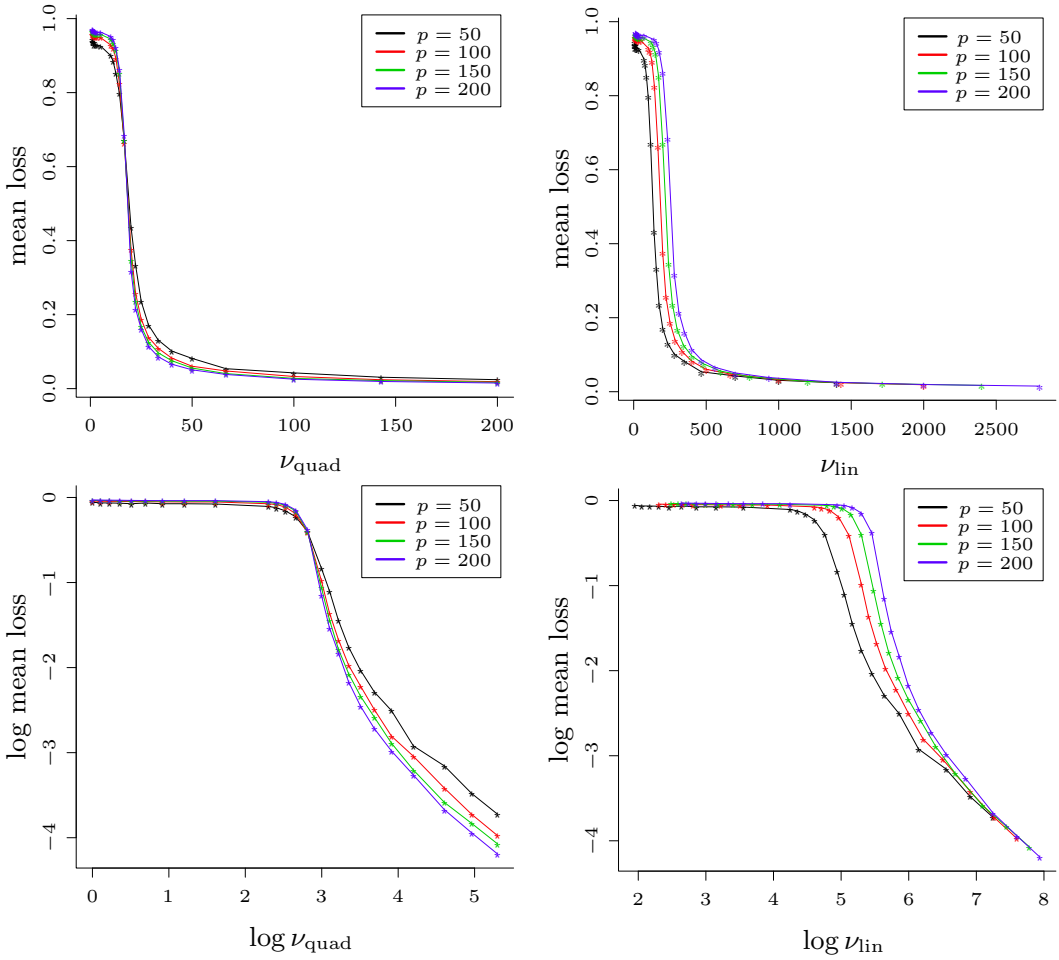

Figure 2.1: Average loss of the estimator $\hat{v}^{\text{SDP}}$ over $N_{\text{rep}} = 100$ repetitions against effective sample sizes ν_{quad} (top left) and ν_{lin} (top right). The tail behaviour under both scalings is examined under logarithmic scales in the bottom left and bottom right panels.

where the final inequality follows because the function $x \mapsto 2e^{-x} + x^m/m!$ is decreasing on $[0, 1/2]$. This calculation allows us to apply Bernstein's inequality (e.g. van de Geer, 2000, Lemma 5.7, taking $K = 2\sigma^2, R = 4\sigma^2$ in her notation), to deduce that for any $s \geq 0$,

$$\mathbb{P}(|\hat{V}(u) - V(u)| \geq s) \leq 2\exp\biggl(-\frac{ns^2}{4\sigma^2 s + 32\sigma^4}\biggr).$$

It follows by Lemma 2.9 in Section 2.7, taking $\epsilon = 1/4$ in that result, that if $\eta > 0$ is such that $\ell \log(p/\eta) \leq n$, then for $C := 8\sigma^2$, we have

$$\begin{aligned}
&\mathbb{P}\biggl(\sup_{u \in \mathbb{S}^{p-1}(\ell)} |\hat{V}(u) - V(u)| \geq 2C\sqrt{\frac{\ell \log(p/\eta)}{n}}\biggr) \\
&\qquad \leq 2\pi\ell^{1/2}\binom{p}{\ell}\Bigl(\frac{128}{\sqrt{255}}\Bigr)^{\ell-1}\exp\biggl(-\frac{C^2\ell\log(p/\eta)}{4C\sigma^2\sqrt{\frac{\ell\log(p/\eta)}{n}} + 32\sigma^4}\biggr) \\
&\qquad \leq 2\pi\ell^{1/2}\Bigl(\frac{e}{\ell}\Bigr)^{\ell}\Bigl(\frac{128}{\sqrt{255}}\Bigr)^{\ell-1}\eta^{\ell} \leq e^9\eta,
\end{aligned}$$

Similarly, if $\ell \log(p/\eta) > n$, then

$$\mathbb{P}\biggl(\sup_{u \in \mathbb{S}^{p-1}(\ell)} |\hat{V}(u) - V(u)| \geq 2C\frac{\ell\log(p/\eta)}{n}\biggr) \leq 2\pi\ell^{1/2}\binom{p}{\ell}\Bigl(\frac{128}{\sqrt{255}}\Bigr)^{\ell-1}\eta^{\ell^2} \leq e^9\eta.$$

Setting $\delta := e^9\eta$, we find (noting that we only need to consider the case $\delta \in (0, 1]$) that

$$\begin{aligned}
&\mathbb{P}\biggl\{\sup_{u \in \mathbb{S}^{p-1}(\ell)} |\hat{V}(u) - V(u)| \geq 16\sigma^2\Bigl(1 + \frac{9}{\log p}\Bigr)\max\biggl(\sqrt{\frac{\ell\log(p/\delta)}{n}}, \frac{\ell\log(p/\delta)}{n}\biggr)\biggr\} \\
&\quad \leq \mathbb{P}\biggl\{\sup_{u \in \mathbb{S}^{p-1}(\ell)} |\hat{V}(u) - V(u)| \geq 16\sigma^2\max\biggl(\sqrt{\frac{\ell\log(e^9p/\delta)}{n}}, \frac{\ell\log(e^9p/\delta)}{n}\biggr)\biggr\} \leq \delta.
\end{aligned}$$

(ii) By Lemma 1 of Laurent and Massart (2000), if $Y_1, \ldots, Y_n$ are independent χ^2_1 random variables, then for all $a > 0$,

$$\mathbb{P}\biggl(\frac{1}{n}\Bigl|\sum_{i=1}^n Y_i - n\Bigr| \geq a\biggr) \leq 2e^{-\frac{n}{2}(1+a-\sqrt{1+2a})} \leq 2e^{-n\min(\frac{a}{4}, \frac{a^2}{16})}.$$

Setting $\eta := e^{-n\min(\frac{a}{4}, \frac{a^2}{16})}$, we deduce that

$$\mathbb{P}\biggl\{\frac{1}{n}\Bigl|\sum_{i=1}^n Y_i - n\Bigr| \geq 4\max\biggl(\sqrt{\frac{\log(1/\eta)}{n}}, \frac{\log(1/\eta)}{n}\biggr)\biggr\} \leq 2\eta.$$

Hence, using Lemma 2.9 again, and by a similar calculation to Part (i),

$$\mathbb{P}\biggl\{\sup_{u \in \mathbb{S}^{p-1}(\ell)} |\hat{V}(u) - V(u)| \geq 8\lambda_1(P)\max\biggl(\sqrt{\frac{\log(1/\eta)}{n}}, \frac{\log(1/\eta)}{n}\biggr)\biggr\} \leq e^9p^{\ell}\eta.$$

The result follows on setting $\delta := e^9p^{\ell}\eta$. □

Proof of Theorem 2.2. Fix an arbitrary $P \in \mathcal{P}_p(n, k, \theta)$. For notational simplicity, we write $v := v_1(P)$ and $\hat{v} := \hat{v}^k_{\max}(\hat{\Sigma})$ in this proof. We now exploit the Curvature Lemma of Vu et al. (2013,

Lemma 3.1), which is closely related to the Davis–Kahan $\sin\theta$ theorem (cf. Davis and Kahan (1970) and Chapter 1. This lemma gives that

$$\|\hat{v}\hat{v}^\top - vv^\top\|_2^2 \leq \frac{2}{\theta}\mathrm{tr}\big(\Sigma(vv^\top - \hat{v}\hat{v}^\top)\big) \leq \frac{2}{\theta}\mathrm{tr}\big((\Sigma - \hat{\Sigma})(vv^\top - \hat{v}\hat{v}^\top)\big).$$

When $\hat{v}\hat{v}^\top \neq vv^\top$, we have that $\frac{vv^\top - \hat{v}\hat{v}^\top}{\|vv^\top - \hat{v}\hat{v}^\top\|_2}$ has rank 2, trace 0 and has non-zero entries in at most $2k$ rows and $2k$ columns. It follows that its non-zero eigenvalues are $\pm 1/\sqrt{2}$, so it can be written as $(xx^\top - yy^\top)/\sqrt{2}$ for some $x, y \in \mathbb{S}^{p-1}(2k)$. Thus

$$\begin{aligned}\mathbb{E}L(\hat{v}, v) = \mathbb{E}\frac{1}{\sqrt{2}}\|\hat{v}\hat{v}^\top - vv^\top\|_2 &\leq \frac{1}{\theta}\mathbb{E}\mathrm{tr}\big((\Sigma - \hat{\Sigma})(xx^\top - yy^\top)\big)\\ &\leq \frac{2}{\theta}\mathbb{E}\sup_{u\in\mathbb{S}^{p-1}(2k)}|\hat{V}(u) - V(u)| \leq 2\sqrt{2}\bigg(1 + \frac{1}{\log p}\bigg)\sqrt{\frac{k\log p}{n\theta^2}},\end{aligned}$$

where we have used Proposition 2.8 in Section 2.7 to obtain the final inequality. □

Proof of Theorem 2.3. Set $\sigma^2 := \frac{1}{8(1+\frac{9}{\log p})} - \theta$. By Proposition 2.1(ii), we have that $N_p(0, \sigma^2 I_p + \theta v_1 v_1^\top)$ belongs to $\mathcal{P}_p(n, k, \theta)$ for any unit vector $v_1 \in \mathbb{S}^{p-1}(k)$. Define $k_0 := k - 1$ and $p_0 := p - 1$. Applying the variant of the Gilbert–Varshamov lemma given as Lemma 2.10 in Section 2.7 with $\alpha := 1/2$ and $\beta := 1/4$, we can construct a set $\mathcal{N}_0$ of k_0-sparse vectors in $\{0,1\}^{p_0}$ with cardinality at least $(p_0/k_0)^{k_0/8}$, such that the Hamming distance between every pair of distinct points in $\mathcal{N}_0$ is at least k_0. For $\epsilon \in (0, 1]$ to be chosen later, define a set of k-sparse vectors in $\mathbb{R}^p$ by

$$\mathcal{N} := \bigg\{\begin{pmatrix}\sqrt{1-\epsilon^2}\\ k_0^{-1/2}\epsilon u_0\end{pmatrix} : u_0 \in \mathcal{N}_0\bigg\}.$$

Observe that if u, v are distinct elements of $\mathcal{N}$, then

$$L(u, v) = \{1 - (u^\top v)^2\}^{1/2} \geq \{1 - (1 - \epsilon^2/2)^2\}^{1/2} \geq \sqrt{3}\epsilon/2,$$

and similarly $L(u, v) \leq \epsilon$. For $u \in \mathcal{N}$, let P_u denote the multivariate normal distribution $N_p(0, \sigma^2 I_p + \theta uu^\top)$. For any estimator $\hat{v} \in \mathcal{V}_{n,p}$, we define $\hat{\psi}_{\hat{v}} := \mathrm{sargmin}_{u\in\mathcal{N}} L(\hat{v}, u)$, where sargmin denotes the smallest element of the argmin in the lexicographic ordering. Note that $\{\hat{\psi}_{\hat{v}} \neq u\} \subseteq \{L(\hat{v}, u) \geq \sqrt{3}\epsilon/4\}$. We now apply the generalised version of Fano's lemma given as Lemma 2.11 in Section 2.7. Writing $D(P\|Q)$ for the Kullback–Leibler divergence between two probability measures defined on the same space, we have

$$\begin{aligned}\inf_{\hat{v}\in\mathcal{V}_{n,p}}\sup_{P\in\mathcal{P}_p(n,k,\theta)}\mathbb{E}_P L\big(\hat{v}, v_1(P)\big) &\geq \inf_{\hat{v}\in\mathcal{V}_{n,p}}\max_{u\in\mathcal{N}}\mathbb{E}_{P_u}L(\hat{v}, u)\\ &\geq \frac{\sqrt{3}\epsilon}{4}\inf_{\hat{v}\in\mathcal{V}_{n,p}}\max_{u\in\mathcal{N}}P_u^{\otimes n}(\hat{\psi}_{\hat{v}} \neq u) \geq \frac{\sqrt{3}\epsilon}{4}\bigg(1 - \frac{\max_{u,v\in\mathcal{N},u\neq v} D(P_v^{\otimes n}\|P_u^{\otimes n}) + \log 2}{(k_0/8)\log(p_0/k_0)}\bigg). \quad (2.8)\end{aligned}$$

We can compute, for distinct points $u, v \in \mathcal{N}$,

$$\begin{aligned}D(P_v^{\otimes n}\|P_u^{\otimes n}) = nD(P_v\|P_u) &= \frac{n}{2}\mathrm{tr}\big((\sigma^2 I_p + \theta uu^\top)^{-1}(\sigma^2 I_p + \theta vv^\top) - I_p\big)\\ &= \frac{n}{2}\mathrm{tr}\big((\sigma^2 I_p + \theta uu^\top)^{-1}\theta(vv^\top - uu^\top)\big) = \frac{n\theta^2 L^2(u, v)}{2\sigma^2(\sigma^2 + \theta)} \leq \frac{n\theta^2\epsilon^2}{2\sigma^2(\sigma^2 + \theta)}. \quad (2.9)\end{aligned}$$

Let $\epsilon := \min\{\sqrt{a/(3b)}, 1\}$, where $a := 1 - \frac{8\log 2}{k_0 \log(p_0/k_0)}$ and $b := \frac{4n\theta^2}{\sigma^2(\sigma^2+\theta)k_0\log(p_0/k_0)}$. Then from (2.8) and (2.9), we find that

$$\inf_{\hat{v}\in\mathcal{V}_{n,p}} \sup_{P\in\mathcal{P}_p(n,k,\theta)} \mathbb{E}_P L\big(\hat{v}, v_1(P)\big) \geq \min\bigg\{\frac{1}{1660}\sqrt{\frac{k\log p}{n\theta^2}}, \frac{5}{18\sqrt{3}}\bigg\},$$

as required. □

Proof of Lemma 2.4. For convenience, we write $v := v_1(\Sigma)$, $\hat{v} := \hat{v}^{\mathrm{SDP}}$ and $\hat{M} := \hat{M}^{\epsilon}$ in this proof. We first study $vv^\top - \hat{M}$, where $\hat{M} \in \mathcal{M}_1$ is computed in Step 2 of Algorithm 2.1. By the Curvature Lemma of Vu et al. (2013, Lemma 3.1),

$$\|vv^\top - \hat{M}\|_2^2 \leq \frac{2}{\theta}\mathrm{tr}\big(\Sigma(vv^\top - \hat{M})\big).$$

Moreover, since $vv^\top \in \mathcal{M}_1$, we have the basic inequality

$$\mathrm{tr}(\hat{\Sigma}\hat{M}) - \lambda\|\hat{M}\|_1 \geq \mathrm{tr}(\hat{\Sigma}vv^\top) - \lambda\|vv^\top\|_1 - \epsilon.$$

Let S denote the set of indices corresponding to the non-zero components of v, and recall that $|S| \leq k$. Since by hypothesis $\|\hat{\Sigma} - \Sigma\|_\infty \leq \lambda$, we have

$$\begin{aligned}
\|vv^\top - \hat{M}\|_2^2 &\leq \frac{2}{\theta}\big\{\mathrm{tr}\big(\hat{\Sigma}(vv^\top - \hat{M})\big) + \mathrm{tr}\big((\Sigma - \hat{\Sigma})(vv^\top - \hat{M})\big)\big\} \\
&\leq \frac{2}{\theta}\big(\lambda\|vv^\top\|_1 - \lambda\|\hat{M}\|_1 + \epsilon + \|\hat{\Sigma} - \Sigma\|_\infty\|vv^\top - \hat{M}\|_1\big) \\
&\leq \frac{2\lambda}{\theta}\big(\|v_S v_S^\top\|_1 - \|\hat{M}_{S,S}\|_1 + \|v_S v_S^\top - \hat{M}_{S,S}\|_1\big) + \frac{2\epsilon}{\theta} \\
&\leq \frac{4\lambda}{\theta}\|v_S v_S^\top - \hat{M}_{S,S}\|_1 + \frac{2\epsilon}{\theta} \leq \frac{4\lambda k}{\theta}\|vv^\top - \hat{M}\|_2 + \frac{2\epsilon}{\theta}.
\end{aligned}$$

We deduce that

$$\|vv^\top - \hat{M}\|_2 \leq \frac{4\lambda k}{\theta} + \sqrt{\frac{2\epsilon}{\theta}}.$$

On the other hand,

$$\begin{aligned}
\|vv^\top - \hat{M}\|_2^2 = \mathrm{tr}\big((vv^\top - \hat{M})^2\big) &= 1 - 2v^\top\hat{M}v + \mathrm{tr}(\hat{M}^2) \\
&\geq 1 - 2\hat{v}^\top\hat{M}\hat{v} + \mathrm{tr}(\hat{M}^2) = \|\hat{v}\hat{v}^\top - \hat{M}\|_2^2.
\end{aligned}$$

We conclude that

$$L(\hat{v}, v) = \frac{1}{\sqrt{2}}\|\hat{v}\hat{v}^\top - vv^\top\|_2 \leq \frac{1}{\sqrt{2}}(\|\hat{v}\hat{v}^\top - \hat{M}\|_2 + \|vv^\top - \hat{M}\|_2) \leq \sqrt{2}\|vv^\top - \hat{M}\|_2 \leq \frac{4\sqrt{2}\lambda k}{\theta} + 2\sqrt{\frac{\epsilon}{\theta}},$$

as required. □

Proof of Theorem 2.5. Fix $P \in \mathcal{P}_p(n, k, \theta)$. By Lemma 2.4, and by Lemma 2.12 in Section 2.7,

$$\begin{aligned}
\mathbb{E}L\big(\hat{v}^{\mathrm{SDP}}, v_1(P)\big) &= \mathbb{E}\big\{L\big(\hat{v}^{\mathrm{SDP}}, v_1(P)\big)\mathbb{1}_{\{\|\hat{\Sigma}-\Sigma\|_\infty \leq \lambda\}}\big\} + \mathbb{E}\big\{L\big(\hat{v}^{\mathrm{SDP}}, v_1(P)\big)\mathbb{1}_{\{\|\hat{\Sigma}-\Sigma\|_\infty > \lambda\}}\big\} \\
&\leq \frac{4\sqrt{2}\lambda k}{\theta} + 2\sqrt{\frac{\epsilon}{\theta}} + \mathbb{P}\bigg(\sup_{u\in\mathbb{S}^{p-1}(2)}\big|\hat{V}(u) - V(u)\big| > 2\sqrt{\frac{\log p}{n}}\bigg) \qquad (2.10)
\end{aligned}$$

Since $P \in \mathrm{RCC}_p(n,2,1)$, we have for each $\delta > 0$ that

$$\mathbb{P}\Big\{ \sup_{u \in \mathbb{S}^{p-1}(2)} \big|\hat{V}(u) - V(u)\big| > \max\Big(\sqrt{\frac{2\log(p/\delta)}{n}}, \frac{2\log(p/\delta)}{n} \Big) \Big\} \leq \delta.$$

Set $\delta := \sqrt{\frac{k^2 \log p}{n\theta^2}}$. Since $4\log p \leq n$, which in particular implies $n \geq 3$, we have

$$\frac{2\log(p/\delta)}{n} \leq \frac{1}{2} + \frac{1}{n}\log\Big(\frac{n\theta^2}{k^2 \log p}\Big) \leq \frac{1}{2} + \frac{\log n}{n} - \frac{1}{n}\log\log 2 \leq 1.$$

Moreover, since $n \leq k^2 p^2 \theta^{-2} \log p$,

$$2\log(p/\delta) = 2\log p + \log\Big(\frac{n\theta^2}{k^2 \log p}\Big) \leq 4 \log p.$$

We deduce that

$$\mathbb{P}\Big(\sup_{u \in \mathbb{S}^{p-1}(2)} \big|\hat{V}(u) - V(u)\big| > 2\sqrt{\frac{\log p}{n}} \Big) \leq \sqrt{\frac{k^2 \log p}{n\theta^2}}. \tag{2.11}$$

The desired risk bound follows from (2.10), the fact that $\theta \leq k$, and (2.11). □

Proof of Theorem 2.6. Suppose, for a contradiction, that there exist an infinite subset $\mathcal{N}$ of $\mathbb{N}$, $K_0 \in [0,\infty)$ and a sequence $(\hat{v}^{(n)})$ of randomised polynomial time estimators of $v_1(P)$ satisfying

$$\sup_{P \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_P L\big(\hat{v}^{(n)}(\mathbf{X}), v_1(P)\big) \leq K_0 \sqrt{\frac{k^{1+\alpha}\log p}{n\theta^2}}$$

for all $n \in \mathcal{N}$. Let $L := \lceil \log p_n \rceil$, let $m = m_n := \lceil 10 L p_n / 9 \rceil$ and let $\kappa = \kappa_n := L k_n$. We claim that Algorithm 2.4 below is a randomised polynomial time algorithm that correctly identifies the Planted Clique problem on m_n vertices and a planted clique of size κ_n with probability tending to 1 as $n \to \infty$. Since $\kappa_n = O(m_n^{1/2-\tau-\delta}\log m_n)$, this contradicts Assumption (A1)(τ). We prove the claim below.

Let $G \sim \mathbb{G}_{m,\kappa}$, and let $K \subseteq V(G)$ denote the planted clique. Note that the matrix $\mathbf{A}$ defined in Step 1 of Algorithm 2.4 is the off-diagonal block of the adjacency matrix of G associated with the bipartite graph induced by the two parts $\{u_i : i = 1,\dots,n\}$ and $\{w_j : j = 1,\dots,p\}$. Let $\boldsymbol{\epsilon'} = (\epsilon'_1,\dots,\epsilon'_n)^\top$ and $\boldsymbol{\gamma'} = (\gamma'_1,\dots,\gamma'_p)^\top$, where $\epsilon'_i := \mathbb{1}_{\{u_i \in K\}}$, $\gamma'_j := \mathbb{1}_{\{w_j \in K\}}$, and set $S' := \{j : \gamma'_j = 1\}$.

It is convenient at this point to introduce the notion of a *Graph Vector distribution*. We say Y has a p-variate Graph Vector distribution with parameters $g = (g_1,\dots,g_p)^\top \in \{0,1\}^p$ and $\pi_0 \in [0,1]$, and write $Y \sim \mathrm{GV}_p^g(\pi_0)$, if we can write

$$Y = \xi\big\{(1-\epsilon)R + \epsilon(g + \tilde{R})\big\},$$

where ξ, ϵ and R are independent, where ξ is a Rademacher random variable, where $\epsilon \sim \mathrm{Bern}(\pi_0)$, where $R = (R_1,\dots,R_p)^\top \in \mathbb{R}^p$ has independent Rademacher components, and where $\tilde{R} = (\tilde{R}_1,\dots,\tilde{R}_p)^\top$ with $\tilde{R}_j := (1-g_j)R_j$.

Let $(\boldsymbol{\epsilon},\boldsymbol{\gamma})^\top = (\epsilon_1,\dots,\epsilon_n,\gamma_1,\dots,\gamma_p)^\top$ be $n+p$ independent $\mathrm{Bern}(\kappa/m)$ random variables. For $i = 1,\dots,n$, let $Y_i := \xi_i\big\{(1-\epsilon_i)R_i + \epsilon_i(\boldsymbol{\gamma} + \tilde{R}_i)\big\}$ so that, conditional on $\boldsymbol{\gamma}$, the random vectors $Y_1,\dots,Y_n$ are independent, each distributed as $\mathrm{GV}_p^{\boldsymbol{\gamma}}(\kappa/m)$. As shorthand, we denote this condi-

Algorithm 2.4: Pseudo-code for a planted clique algorithm based on a hypothetical randomised polynomial time sparse principal component estimation algorithm.

Input: $m \in \mathbb{N}$, $\kappa \in \{1, \ldots, m\}$, $G \in \mathbb{G}_m$, $L \in \mathbb{N}$

begin

Step 1: Let $n \leftarrow \lfloor 9m/(10L) \rfloor$, $p \leftarrow p_n$, $k \leftarrow \lfloor \kappa/L \rfloor$. Draw $u_1, \ldots, u_n, w_1, \ldots, w_p$ uniformly at random without replacement from $V(G)$. Form $\mathbf{A} = (A_{ij}) \leftarrow (\mathbb{1}_{\{u_i \sim w_j\}}) \in \mathbb{R}^{n \times p}$ and $\mathbf{X} \leftarrow \mathrm{diag}(\xi_1, \ldots, \xi_n)(2\mathbf{A} - \mathbf{1}_{n \times p})$, where $\xi_1, \ldots, \xi_n$ are independent Rademacher random variables (independent of $u_1, \ldots, u_n, w_1, \ldots, w_p$), and where every entry of $\mathbf{1}_{n \times p} \in \mathbb{R}^{n \times p}$ is 1.

Step 2: Use the randomised estimator $\hat{v}^{(n)}$ to compute $\hat{v} = \hat{v}^{(n)}(\mathbf{X}/\sqrt{750})$.

Step 3: Let $\hat{S} = \hat{S}(\hat{v})$ be the lexicographically smallest k-subset of $\{1, \ldots, p\}$ such that $(\hat{v}_j : j \in \hat{S})$ contains the k largest coordinates of $\hat{v}$ in absolute value.

Step 4: For $u \in V(G)$ and $W \subseteq V(G)$, let $\mathrm{nb}(u, W) := \mathbb{1}_{\{u \in W\}} + \sum_{w \in W} \mathbb{1}_{\{u \sim w\}}$. Set $\hat{K} := \big\{u \in V(G) : \mathrm{nb}(u, \{w_j : j \in \hat{S}\}) \geq 3k/4\big\}$.

end

Output: $\hat{K}$

tional distribution as $Q_{\boldsymbol{\gamma}}$, and write $S := \{j : \gamma_j = 1\}$. Note that by Lemma 2.13 in Section 2.7, $Q_{\boldsymbol{\gamma}} \in \cap_{\ell=1}^{\lfloor 20p/(9k) \rfloor} \mathrm{RCC}_p(\ell, 750)$.

Let $\boldsymbol{Y} := (Y_1, \ldots, Y_n)^\top$. Recall that if P and Q are probability measures on a measurable space $(\mathcal{X}, \mathcal{B})$, the *total variation distance* between P and Q is defined by

$$d_{\mathrm{TV}}(P, Q) := \sup_{B \in \mathcal{B}} |P(B) - Q(B)|.$$

Writing $\mathcal{L}(Z)$ for the distribution (or law) of a generic random element Z, and using elementary properties of the total variation distance given in Lemma 2.16 in Section 2.7, we have

$$\begin{aligned} d_{\mathrm{TV}}\big(\mathcal{L}(\mathbf{X}), \mathcal{L}(\boldsymbol{Y})\big) &\leq d_{\mathrm{TV}}\Big(\mathcal{L}\big(\boldsymbol{\epsilon}', \boldsymbol{\gamma}', (R_{ij}), (\xi_i)\big), \mathcal{L}\big(\boldsymbol{\epsilon}, \boldsymbol{\gamma}, (R_{ij}), (\xi_i)\big)\Big) \\ &= d_{\mathrm{TV}}\big(\mathcal{L}(\boldsymbol{\epsilon}', \boldsymbol{\gamma}'), \mathcal{L}(\boldsymbol{\epsilon}, \boldsymbol{\gamma})\big) \leq \frac{2(n+p)}{m} \leq \frac{9(n+p)}{5p \log p}. \end{aligned} \tag{2.12}$$

Here, the penultimate inequality follows from Diaconis and Freedman (1980, Theorem 4). In view of (2.12), we initially analyse Steps 2, 3 and 4 in Algorithm 2.4 with $\mathbf{X}$ replaced by $\boldsymbol{Y}$. Observe that $\mathbb{E}(Y_i|\boldsymbol{\gamma}) = 0$ and, writing $\Delta := \mathrm{diag}(\boldsymbol{\gamma}) \in \mathbb{R}^{p \times p}$, we have

$$\begin{aligned} \Sigma_{\boldsymbol{\gamma}} := \mathrm{Cov}(Y_i|\boldsymbol{\gamma}) &= \mathbb{E}\big\{(1 - \epsilon_i) R_i R_i^\top + \epsilon_i (\boldsymbol{\gamma} + \tilde{R}_i)(\boldsymbol{\gamma} + \tilde{R}_i)^\top \big| \boldsymbol{\gamma}\big\} \\ &= I_p + \frac{\kappa}{m}(\boldsymbol{\gamma}\boldsymbol{\gamma}^\top - \Delta). \end{aligned}$$

Writing $N_{\boldsymbol{\gamma}} := \sum_{j=1}^p \gamma_j$, it follows that the largest eigenvalue of $\Sigma_{\boldsymbol{\gamma}}$ is $1 + \frac{\kappa}{m}(N_{\boldsymbol{\gamma}} - 1)$, with corresponding eigenvector $\boldsymbol{\gamma}/N_{\boldsymbol{\gamma}}^{1/2} \in \mathbb{S}^{p-1}(N_{\boldsymbol{\gamma}})$. The other eigenvalues are 1, with multiplicity $p - N_{\boldsymbol{\gamma}}$, and $1 - \frac{\kappa}{m}$, with multiplicity $N_{\boldsymbol{\gamma}} - 1$. Hence $\lambda_1(\Sigma_{\boldsymbol{\gamma}}) - \lambda_2(\Sigma_{\boldsymbol{\gamma}}) = \frac{\kappa}{m}(N_{\boldsymbol{\gamma}} - 1)$. Define

$$\Gamma_0 := \left\{ g \in \{0,1\}^p : \left| N_g - \frac{p\kappa}{m} \right| \leq \frac{k}{20} \right\},$$

where $N_g := \sum_{j=1}^p g_j$. We note that by Bernstein's inequality (e.g. Shorack and Wellner, 1986, p. 855) that

$$\mathbb{P}(\boldsymbol{\gamma} \in \Gamma_0) \geq 1 - 2e^{-k/800}. \tag{2.13}$$

If $g \in \Gamma_0$, the conditional distribution of $Y_1/\sqrt{750}$ given $\boldsymbol{\gamma} = g$ belongs to classes $\mathcal{P}_p(n, k, \theta)$ for $\theta \le \frac{\kappa}{750m}(N_g - 1)$ and all $n \in \mathcal{N}$ sufficiently large. By hypothesis, it follows that for $g \in \Gamma_0$,

$$\mathbb{E}\Big\{ L\big(\hat{v}^{(n)}(\boldsymbol{Y}/\sqrt{750}), v_1(Q_{\boldsymbol{\gamma}})\big) \;\Big|\; \boldsymbol{\gamma} = g \Big\} \le K_0 \sqrt{\frac{k^{1+\alpha} \log p}{n\theta^2}}$$

for all large $n \in \mathcal{N}$. Then by Lemma 2.14 in Section 2.7, for $\hat{S}(\cdot)$ defined in Step 3 of Algorithm 2.4, for $g \in \Gamma_0$, and large $n \in \mathcal{N}$,

$$\begin{aligned}\mathbb{E}\Big\{ \big| S \setminus \hat{S}\big(\hat{v}^{(n)}(\boldsymbol{Y}/\sqrt{750})\big)\big| \;\Big|\; \boldsymbol{\gamma} = g \Big\} &\le 2N_g \mathbb{E}\Big\{ L\big(\hat{v}^{(n)}(\boldsymbol{Y}/\sqrt{750}), v_1(Q_{\boldsymbol{\gamma}})\big)^2 \;\Big|\; \boldsymbol{\gamma} = g \Big\} \\ &\le 2N_g K_0 \sqrt{\frac{k^{1+\alpha} \log p}{n\theta^2}}\end{aligned}$$

We deduce by Markov's inequality that for $g \in \Gamma_0$, and large $n \in \mathcal{N}$,

$$\mathbb{P}\Big\{ \big| S \cap \hat{S}\big(\hat{v}^{(n)}(\boldsymbol{Y}/\sqrt{750})\big)\big| \le 16N_{\boldsymbol{\gamma}}/17 \;\Big|\; \boldsymbol{\gamma} = g \Big\} \le 34K_0 \sqrt{\frac{k^{1+\alpha} \log p}{n\theta^2}} \tag{2.14}$$

Let

$$\begin{aligned}\Omega_{0,n} &:= \{\boldsymbol{\gamma} \in \Gamma_0\} \cap \big\{ \big| S \cap \hat{S}\big(\hat{v}^{(n)}(\boldsymbol{Y}/\sqrt{750})\big)\big| > 16N_{\boldsymbol{\gamma}}/17 \big\} \\ \Omega'_{0,n} &:= \{\boldsymbol{\gamma}' \in \Gamma_0\} \cap \big\{ \big| S \cap \hat{S}\big(\hat{v}^{(n)}(\boldsymbol{X}/\sqrt{750})\big)\big| > 16N_{\boldsymbol{\gamma}'}/17 \big\} =: \Omega'_{1,n} \cap \Omega'_{2,n},\end{aligned}$$

say, where $N_{\boldsymbol{\gamma}'} := \sum_{j=1}^p \gamma'_j$. When $n \in \mathcal{N}$ is sufficiently large, we have on the event $\Omega'_{0,n}$ that

$$\big| \big\{ j \in \hat{S}\big(\hat{v}^{(n)}(\boldsymbol{X}/\sqrt{750})\big) : w_j \in K \big\} \big| > 3k/4. \tag{2.15}$$

Now set

$$\Omega'_{3,n} := \big\{ \mathrm{nb}(u, \{w_j : j \in S'\}) \le k/2 \text{ for all } u \in V(G) \setminus K \big\}.$$

Recall the definition of $\hat{K}$ from Step 4 of Algorithm 2.4. We claim that for sufficiently large $n \in \mathcal{N}$, $\Omega'_{0,n} \cap \Omega'_{3,n} \subseteq \{\hat{K} = K\}$. To see this, note that for $n \in \mathcal{N}$ sufficiently large, on $\Omega'_{0,n}$ we have $K \subseteq \hat{K}$ by (2.15). For the reverse inclusion, note that if $u \in V(G) \setminus K$, then on $\Omega'_{0,n} \cap \Omega'_{3,n}$, we have for sufficiently large $n \in \mathcal{N}$ that

$$\begin{aligned}\mathrm{nb}\Big(u, \big\{ w_j : j \in \hat{S}\big(\hat{v}^{(n)}(\boldsymbol{X}/\sqrt{750})\big)\big\}\Big) &\le \big| \{w_j : j \in \hat{S}\} \setminus K \big| + \mathrm{nb}\big(u, \{w_j : j \in \hat{S}\} \cap K\big) \\ &\le \big| \{w_j : j \in \hat{S}\} \setminus K \big| + \mathrm{nb}\big(u, \{w_j : j \in S'\}\big) < \frac{k}{4} + \frac{k}{2} = \frac{3k}{4}.\end{aligned}$$

This establishes our claim. We conclude that for sufficiently large $n \in \mathcal{N}$,

$$\mathbb{P}(\hat{K} \ne K) \le \mathbb{P}\big((\Omega'_{0,n} \cap \Omega'_{3,n})^c\big) \le \mathbb{P}\big((\Omega'_{0,n})^c\big) + \mathbb{P}\big(\Omega'_{1,n} \cap (\Omega'_{3,n})^c\big). \tag{2.16}$$

Now, by Lemma 2.16, we have

$$|\mathbb{P}(\Omega'_{0,n}) - \mathbb{P}(\Omega_{0,n})| \le d_{\mathrm{TV}}\big(\mathcal{L}(\mathbf{X}, \boldsymbol{\gamma}'), \mathcal{L}(\mathbf{Y}, \boldsymbol{\gamma})\big) \le \frac{9(n+p)}{5p \log p}. \tag{2.17}$$

Moreover, by a union bound and Hoeffding's inequality, for large $n \in \mathcal{N}$,

$$\mathbb{P}\big(\Omega'_{1,n} \cap (\Omega'_{3,n})^c\big) \le \sum_{g \in \Gamma_0} \mathbb{P}\big((\Omega'_{3,n})^c | \boldsymbol{\gamma} = g\big)\mathbb{P}(\boldsymbol{\gamma} = g) \le m e^{-k/800}. \tag{2.18}$$

We conclude by (2.16), (2.17), (2.13), (2.14) and (2.18) that for large $n \in \mathcal{N}$,

$$\mathbb{P}(\hat{K} \neq K) \le \frac{9(n+p)}{5p \log p} + 2e^{-k/800} + 34K_0\sqrt{\frac{k^{1+\alpha}\log p}{n\theta^2}} + me^{-k/800} \to 0$$

as $n \to \infty$. This contradicts Assumption (A1)(τ), and therefore completes the proof. □

Proof of Theorem 2.7. Setting $\delta := p^{-1}$ in (2.3), there exist events Ω_1 and Ω_2, each with probability at least $1 - p^{-1}$, such that on Ω_1 and Ω_2, we respectively have

$$\sup_{u \in \mathbb{S}^{p-1}(2k)} |\hat{V}(u) - V(u)| \le 2\sqrt{\frac{k \log p}{n}} \quad \text{and} \quad \sup_{u \in \mathbb{S}^{p-1}(2)} |\hat{V}(u) - V(u)| \le 2\sqrt{\frac{\log p}{n}}. \tag{2.19}$$

Let $\Omega_0 := \Omega_1 \cap \Omega_2$. We work on Ω_0 henceforth. The main ingredient for proving both parts of the theorem is the following weak-duality inequality:

$$\begin{aligned}\max_{M \in \mathcal{M}_1} \operatorname{tr}(\hat{\Sigma}M) - \lambda\|M\|_1 &= \max_{M \in \mathcal{M}_1} \min_{U \in \mathcal{U}} \operatorname{tr}\big((\hat{\Sigma} - U)M\big) \\ &\le \min_{U \in \mathcal{U}} \max_{M \in \mathcal{M}_1} \operatorname{tr}\big((\hat{\Sigma} - U)M\big) = \min_{U \in \mathcal{U}} \lambda_1(\hat{\Sigma} - U). \end{aligned} \tag{2.20}$$

It is convenient to denote $\gamma := \sqrt{\frac{k^2 \log p}{n\theta^2}}$, and note that

$$\gamma \le \frac{\sqrt{k}}{16} \min_{j \in S} |v_{1,j}| \le \frac{1}{16}\|v_{1,S}\|_2 = \frac{1}{16}.$$

Proof of (a). From (2.20), it suffices to exhibit a primal-dual pair $(\hat{M}, \hat{U}) \in \mathcal{M}_1 \times \mathcal{U}$, such that

(C1) $\hat{M} = \hat{v}\hat{v}^\top$ with $\operatorname{sgn}(\hat{v}) = \operatorname{sgn}(v_1)$.

(C2) $\operatorname{tr}(\hat{\Sigma}\hat{M}) - \lambda\|\hat{M}\|_1 = \lambda_1(\hat{\Sigma} - \hat{U})$.

We construct the primal-dual pair as follows. Define

$$\hat{U} := \begin{pmatrix} \lambda \operatorname{sgn}(v_{1,S})\operatorname{sgn}(v_{1,S})^\top & \hat{\Sigma}_{SS^c} - \Sigma_{SS^c} \\ \hat{\Sigma}_{S^cS} - \Sigma_{S^cS} & \hat{\Sigma}_{S^cS^c} - \Sigma_{S^cS^c} \end{pmatrix}.$$

By (2.19) and Lemma 2.12, we have that $\|\hat{\Sigma} - \Sigma\|_\infty \le 4\sqrt{\frac{\log p}{n}} \le \lambda$, so $U \in \mathcal{U}$. Let $w = (w_1, \ldots, w_k)$ be a unit-length leading eigenvector of $\Sigma_{SS} - \hat{U}_{SS}$ such that $w^\top v_{1,S} \ge 0$. Then, define $\hat{v}$ componentwise by

$$\hat{v}_S \in \operatorname*{argmax}_{\substack{u \in \mathbb{R}^k, \|u\|_2 = 1 \\ u^\top w \ge 0}} u^\top \Big(\hat{\Sigma}_{SS} - \hat{U}_{SS}\Big)u, \quad \hat{v}_{S^c} = 0,$$

and set $\hat{M} := \hat{v}\hat{v}^\top$. Note that our choices above ensure that $\hat{M} \in \mathcal{M}_1$. To verify (C1), we now show that $\operatorname{sgn}(\hat{v}_S) = \operatorname{sgn}(w) = \operatorname{sgn}(v_{1,S})$. By a variant of the Davis–Kahan theorem (Yu, Wang

and Samworth, 2015, Theorem 2),

$$\|w-\hat{v}_S\|_\infty \le \|w-\hat{v}_S\|_2 \le \sqrt{2}L(\hat{v}_S,w) \le \frac{2\sqrt{2}\|\hat{\Sigma}_{SS}-\Sigma_{SS}\|_{\mathrm{op}}}{\theta}$$
$$\le \frac{2\sqrt{2}}{\theta}\sup_{u\in\mathbb{S}^{p-1}(2k)}|\hat{V}(u)-V(u)| \le 4\sqrt{2}\gamma k^{-1/2}, \tag{2.21}$$

where the final inequality uses (2.19). But w is also a leading eigenvector of

$$\frac{1}{\theta}(\Sigma_{SS}-\hat{U}_{SS}-I_k) = v_{1,S}v_{1,S}^\top - 4\gamma ss^\top,$$

where $s := \frac{\mathrm{sgn}(v_{1,S})}{\|\mathrm{sgn}(v_{1,S})\|}$. Write $s = \alpha v_{1,S}+\beta v_\perp$ for some $\alpha,\beta\in\mathbb{R}$ with $\alpha^2+\beta^2=1$, and a unit vector $v_\perp\in\mathbb{R}^k$ orthogonal to $v_{1,S}$. Then

$$v_{1,S}v_{1,S}^\top - 4\gamma ss^\top = \begin{pmatrix} v_{1,S} & v_\perp\end{pmatrix}\begin{pmatrix}1-4\gamma\alpha^2 & -4\gamma\alpha\beta\\ -4\gamma\alpha\beta & -4\gamma\beta^2\end{pmatrix}\begin{pmatrix}v_{1,S}^\top\\ v_\perp^\top\end{pmatrix}$$
$$= \begin{pmatrix} v_{1,S} & v_\perp\end{pmatrix}\begin{pmatrix}a_1 & b_1\\ a_2 & b_2\end{pmatrix}\begin{pmatrix}d_1 & 0\\ 0 & d_2\end{pmatrix}\begin{pmatrix}a_1 & a_2\\ b_1 & b_2\end{pmatrix}\begin{pmatrix}v_{1,S}^\top\\ v_\perp^\top\end{pmatrix},$$

where $d_1\ge d_2$ and $\begin{pmatrix}a_1 & a_2\end{pmatrix}^\top$, $\begin{pmatrix}b_1 & b_2\end{pmatrix}^\top$ are eigenvalues and corresponding unit-length eigenvectors of the middle matrix on the right-hand side of the first line. Direct computation yields that $d_1\ge 1/2>0\ge d_2$ and

$$\begin{pmatrix}a_1\\ a_2\end{pmatrix} \propto \begin{pmatrix}1-4\gamma\alpha^2+4\gamma\beta^2+\sqrt{16\gamma\beta^2+(1-4\gamma)^2}\\ -8\gamma\alpha\beta\end{pmatrix}.$$

Consequently, w is a scalar multiple of

$$a_1v_{1,S}+a_2v_\perp = \Big\{1+4\gamma+\sqrt{16\gamma\beta^2+(1-4\gamma)^2}\Big\}v_{1,S} - 8\gamma\alpha s. \tag{2.22}$$

Since

$$\Big\{1+4\gamma+\sqrt{16\gamma\beta^2+(1-4\gamma)^2}\Big\}\min_{j\in S}|v_{1,j}| \ge 2\min_{j\in S}|v_{1,j}| \ge 32\gamma k^{-1/2} > 8\gamma\alpha\|s\|_\infty,$$

we have $\mathrm{sgn}(w)=\mathrm{sgn}(v_{1,S})$. Hence by (2.22),

$$\min_{j=1,\dots,k}|w_j| \ge \frac{\Big\{1+4\gamma+\sqrt{16\gamma\beta^2+(1-4\gamma)^2}\Big\}\min_{j\in S}|v_{1,j}| - 8\gamma\alpha\|s\|_\infty}{\|a_1v_{1,S}+a_2v_\perp\|_2}$$
$$\ge \frac{(32-8\alpha)\gamma k^{-1/2}}{1+4\gamma+\sqrt{16\gamma\beta^2+(1-4\gamma)^2}} \ge \frac{12\gamma k^{-1/2}}{1+4\gamma} \ge \frac{48}{5}\gamma k^{-1/2}, \tag{2.23}$$

By (2.21) and (2.23), we have $\min_j|w_j| > \|w-\hat{v}_S\|_\infty$. So $\mathrm{sgn}(\hat{v}_S)=\mathrm{sgn}(w)=\mathrm{sgn}(v_{1,S})$ as desired.

It remains to check condition (C2). Since $\mathrm{sgn}(\hat{v}_S)=\mathrm{sgn}(v_{1,S})$, we have

$$\mathrm{tr}(\hat{\Sigma}\hat{M}) - \lambda\|\hat{M}\|_1 = \mathrm{tr}(\hat{\Sigma}_{SS}\hat{v}_S\hat{v}_S^\top) - \mathrm{tr}(\hat{U}_{SS}\hat{v}_S\hat{v}_S^\top)$$
$$= \hat{v}_S^\top(\hat{\Sigma}_{SS}-\hat{U}_{SS})\hat{v}_S = \lambda_1(\hat{\Sigma}_{SS}-\hat{U}_{SS}).$$

Moreover, $\hat{\Sigma} - \hat{U}$ is block diagonal with blocks $\hat{\Sigma}_{SS} - \hat{U}_{SS}$ and Γ_{p-k}. As $\lambda_1(\Gamma_{p-k}) \leq 1$ by assumption, it suffices to show that $\lambda_1(\hat{\Sigma}_{SS} - \hat{U}_{SS}) \geq 1$. By Weyl's inequality (see e.g. Horn and Johnson (2012, Theorem 4.3.1))

$$\begin{aligned}\lambda_1(\hat{\Sigma}_{SS} - \hat{U}_{SS}) &\geq \lambda_1(\Sigma_{SS} - \hat{U}_{SS}) - \|\hat{\Sigma}_{SS} - \Sigma_{SS}\|_{\mathrm{op}} \\ &\geq 1 + \theta\lambda_1(v_{1,S}v_{1,S}^\top - 4\gamma ss^\top) - 2\sqrt{\frac{k\log p}{n}} \geq 1 + \frac{3\theta}{8} > 1. \end{aligned} \tag{2.24}$$

Proof of (b). We claim first that $\hat{S} = S$. Let $\phi^* := f(\hat{M})$ be the optimal value of the semidefinite programme (2.5). From (2.24), we have $\phi^* \geq 1 + 3\theta/8$. The proof strategy here is to use dual matrices $\hat{U}$ defined in part (a) and $\hat{U}'$ to be defined below to respectively bound $\mathrm{tr}(\hat{M}^\epsilon_{S^cS^c})$ from above and bound $\hat{M}^\epsilon_{rr}$ from below for each $r \in S$. We then check that for the choice of ϵ we have in the theorem, the diagonal entries of $\hat{M}^\epsilon$ are above the threshold $\log p/(6n)$ precisely when they belong to the (S, S)-block of the matrix.

From (2.20) and the fact that $\mathrm{tr}(AB) \leq \mathrm{tr}(A)\lambda_1(B)$ for symmetric matrices A and B, we have

$$\begin{aligned}\mathrm{tr}(\hat{\Sigma}\hat{M}^\epsilon) - \lambda\|\hat{M}^\epsilon\|_1 &\leq \mathrm{tr}\big((\hat{\Sigma} - \hat{U})\hat{M}^\epsilon\big) = \mathrm{tr}\big((\hat{\Sigma}_{SS} - \hat{U}_{SS})\hat{M}^\epsilon_{SS}\big) + \mathrm{tr}\big(\Sigma_{S^cS^c}\hat{M}^\epsilon_{S^cS^c}\big) \\ &\leq \mathrm{tr}(\hat{M}^\epsilon_{SS})\phi^* + \mathrm{tr}(\hat{M}^\epsilon_{S^cS^c})\lambda_1(\Gamma_{p-k}) \\ &= \phi^* - \mathrm{tr}(\hat{M}^\epsilon_{S^cS^c})(\phi^* - 1) \leq \phi^* - 3\theta\,\mathrm{tr}(\hat{M}^\epsilon_{S^cS^c})/8.\end{aligned}$$

On the other hand, $\mathrm{tr}(\hat{\Sigma}\hat{M}^\epsilon) - \lambda\|\hat{M}^\epsilon\|_1 \geq \phi^* - \epsilon$. It follows that

$$\mathrm{tr}(\hat{M}^\epsilon_{S^cS^c}) \leq \frac{8\epsilon}{3\theta} \leq \frac{1}{6}\left(\frac{\log p}{Bn}\right)^2 < \tau. \tag{2.25}$$

Next, fix an arbitrary $r \in S$ and define $S_0 := S \setminus \{r\}$. Define $\hat{U}'$ by

$$\hat{U}'_{ij} := \begin{cases} \lambda\,\mathrm{sgn}(\hat{M}_{ij}) & \text{if } i, j \in S_0 \\ \hat{\Sigma}_{ij} - \Sigma_{ij} & \text{otherwise.} \end{cases}$$

We note that on Ω_0, we have $\hat{U}' \in \mathcal{U}$. Again by (2.20),

$$\begin{aligned}\mathrm{tr}(\hat{\Sigma}\hat{M}^\epsilon) - \lambda\|\hat{M}^\epsilon\|_1 &\leq \mathrm{tr}\big((\hat{\Sigma} - \hat{U}')\hat{M}^\epsilon\big) \\ &= \mathrm{tr}\big((\hat{\Sigma}_{S_0S_0} - \hat{U}_{S_0S_0})\hat{M}^\epsilon_{S_0S_0}\big) + \sum_{\substack{(i,j)\in S\times S \\ i=r \text{ or } j=r}} \Sigma_{ij}\hat{M}^\epsilon_{ji} + \mathrm{tr}\big(\Sigma_{S^cS^c}\hat{M}^\epsilon_{S^cS^c}\big) \\ &\leq \mathrm{tr}(\hat{M}^\epsilon_{S_0S_0})\lambda_1\big(\hat{\Sigma}_{S_0S_0} - \hat{U}_{S_0S_0}\big) + \sum_{\substack{(i,j)\in S\times S \\ i=r \text{ or } j=r}} \Sigma_{ij}\hat{M}^\epsilon_{ji} + \mathrm{tr}(\hat{M}^\epsilon_{S^cS^c})\lambda_1(\Gamma_{p-k}).\end{aligned} \tag{2.26}$$

We bound the three terms of (2.26) separately. By Lemma 2.15 in Section 2.7,

$$\lambda_1(\hat{\Sigma}_{S_0S_0} - \hat{U}_{S_0S_0}) \leq \lambda_1(\hat{\Sigma}_{SS} - \hat{U}_{SS}) - \big\{\lambda_1(\hat{\Sigma}_{SS} - \hat{U}_{SS}) - \lambda_2(\hat{\Sigma}_{SS} - \hat{U}_{SS})\big\}\min_{j\in S}\hat{v}_j^2.$$

From (2.21) and (2.23),

$$\min_j |\hat{v}_j| \geq \min_j |w_j| - \|w - \hat{v}_S\|_\infty \geq 3.9\gamma k^{-1/2}.$$

Also, by Weyl's inequality,

$$\begin{aligned}\lambda_1(\hat\Sigma_{SS}-\hat U_{SS})-\lambda_2(\hat\Sigma_{SS}-\hat U_{SS}) &\geq \lambda_1(\Sigma_{SS}-\hat U_{SS})-\lambda_2(\Sigma_{SS}-\hat U_{SS})-2\|\hat\Sigma_{SS}-\Sigma_{SS}\|_{\mathrm{op}}\\ &\geq \theta\Big\{\lambda_1(v_{1,S}v_{1,S}^\top-4\gamma ss^\top)-\lambda_2(v_{1,S}v_{1,S}^\top-4\gamma ss^\top)\Big\}-4\sqrt{\frac{k\log p}{n}}\\ &\geq \theta(1/2-4\gamma k^{-1/2})\geq\theta/4.\end{aligned}$$

It follows that

$$\lambda_1(\hat\Sigma_{S_0S_0}-\hat U_{S_0S_0})\leq\phi^*-3.8\gamma^2k^{-1}\theta. \tag{2.27}$$

For the second term in (2.26), observe that

$$\begin{aligned}\sum_{\substack{(i,j)\in S\times S\\ i=r\,\text{or}\,j=r}}\Sigma_{ij}\hat M^\epsilon_{ij} &\leq (1+\theta v_{1,r}^2)\hat M^\epsilon_{rr}+2\sum_{i\in S,i\neq r}\theta v_{1,i}v_{1,r}\hat M^\epsilon_{i,r}\\ &\leq \hat M^\epsilon_{rr}+2\theta|v_{1,r}|\cdot\|v_1\|_1\sqrt{\hat M^\epsilon_{rr}}\leq\hat M^\epsilon_{rr}+2\theta\sqrt{k}\sqrt{\hat M^\epsilon_{rr}},\end{aligned} \tag{2.28}$$

where the penultimate inequality uses the fact that $\hat M^\epsilon_{ir}\leq\sqrt{\hat M^\epsilon_{ii}\hat M^\epsilon_{rr}}\leq\sqrt{\hat M^\epsilon_{rr}}$ for a non-negative definite matrix $\hat M^\epsilon$. Substituting (2.27) and (2.28) into (2.26),

$$\begin{aligned}\mathrm{tr}(\hat\Sigma\hat M^\epsilon)-\lambda\|\hat M^\epsilon\|_1 &\leq \mathrm{tr}(\hat M^\epsilon_{S_0S_0})\big(\phi^*-\tfrac{3.8\gamma^2\theta}{k}\big)+\hat M^\epsilon_{rr}+2\theta\sqrt{k\hat M^\epsilon_{rr}}+\mathrm{tr}(\hat M^\epsilon_{S^cS^c})\\ &\leq\phi^*-3.8\gamma^2k^{-1}\theta\,\mathrm{tr}(\hat M^\epsilon_{S_0S_0})+2\theta\sqrt{k\hat M^\epsilon_{rr}}\\ &\leq\phi^*-3.8\gamma^2k^{-1}\theta\big\{1-\mathrm{tr}(\hat M^\epsilon_{S^cS^c})\big\}+2\theta(\sqrt{k}+1.9\gamma^2)\sqrt{\hat M^\epsilon_{rr}}.\end{aligned}$$

By definition, $\mathrm{tr}(\hat\Sigma\hat M^\epsilon)-\lambda\|\hat M^\epsilon\|_1\geq\phi^*-\epsilon$, so together with (2.25), we have

$$\begin{aligned}\sqrt{\hat M^\epsilon_{rr}} &\geq \frac{3.8\gamma^2k^{-1}\theta(1-\frac{8\epsilon}{3\theta})-\epsilon}{2\theta(\sqrt{k}+1.9\gamma^2)}\geq\frac{1.9\gamma^2k^{-1}(1-\frac{8\epsilon}{3\theta})}{(\sqrt{k}+\frac{1.9}{256})}-\frac{\epsilon}{2\theta}\\ &\geq 1.8\gamma^2k^{-3/2}\Big(1-\frac{8\epsilon}{3\theta}\Big)-\frac{\epsilon}{2\theta}\\ &\geq\frac{1.8k^{1/2}\log p}{n\theta^2}\left\{1-\frac16\left(\frac{\log p}{Bn}\right)^2\right\}-\frac{1}{32}\left(\frac{\log p}{Bn}\right)^2\geq\frac{1.4\log p}{Bn}>\tau^{1/2}.\end{aligned} \tag{2.29}$$

From (2.25) and (2.29) we conclude that $\hat S=S$, as claimed.

To conclude, by Yu, Wang and Samworth (2015, Theorem 2), on Ω_0,

$$L(\hat v^{\mathrm{MSDP}},v_1)=L(\hat v^{\mathrm{MSDP}}_S,v_{1,S})\leq\frac{2\|\hat\Sigma_{SS}-\Sigma_{SS}\|_{\mathrm{op}}}{\lambda_1(\Sigma_{SS})-\lambda_2(\Sigma_{SS})}\leq 4\sqrt{\frac{k\log p}{n\theta^2}},$$

where we used (2.19) and Lemma 2.12 in the final bound.

For the final part of the theorem, when $p\geq\theta\sqrt{n/k}$,

$$\sup_{P\in\tilde{\mathcal{P}}_p(n,k,\theta)}\mathbb{E}_P\big\{L(\hat v^{\mathrm{MSDP}},v_1)\big\}\leq 4\sqrt{\frac{k\log p}{n\theta^2}}+\mathbb{P}(\Omega_0^c)\leq 4\sqrt{\frac{k\log p}{n\theta^2}}+2p^{-1}\leq 6\sqrt{\frac{k\log p}{n\theta^2}},$$

as desired. □

2.7 Appendix: Ancillary results

We collect here various results used in the proofs in Appendix 2.6.

Proposition 2.8. *Let $P \in \mathrm{RCC}_p(n, \ell, C)$ and suppose that $\ell \log p \le n$. Then*

$$\mathbb{E} \sup_{u \in \mathbb{S}^{p-1}(\ell)} |\hat{V}(u) - V(u)| \le \Big(1 + \frac{1}{\log p}\Big) C \sqrt{\frac{\ell \log p}{n}}.$$

Proof. By setting $\delta = p^{1-t}$ in the RCC condition, we find that

$$\mathbb{P}\bigg(\sup_{u \in \mathbb{S}^{p-1}(\ell)} |\hat{V}(u) - V(u)| \ge C \max\bigg\{ \sqrt{\frac{t\ell \log p}{n}}, \frac{t\ell \log p}{n} \bigg\} \bigg) \le \min(1, p^{1-t})$$

for all $t \ge 0$. It follows that

$$\begin{aligned}
\mathbb{E} \sup_{u \in \mathbb{S}^{p-1}(\ell)} |\hat{V}(u) - V(u)| &= \int_0^\infty \mathbb{P}\bigg(\sup_{u \in \mathbb{S}^{p-1}(\ell)} |\hat{V}(u) - V(u)| \ge s \bigg)\, ds \\
&\le C \sqrt{\frac{\ell \log p}{n}} + C \sqrt{\frac{\ell \log p}{n}} \int_1^{\frac{n}{\ell \log p}} \frac{1}{2} p^{1-t} t^{-1/2}\, dt + C \frac{\ell \log p}{n} \int_{\frac{n}{\ell \log p}}^\infty p^{1-t}\, dt \\
&\le C \sqrt{\frac{\ell \log p}{n}} \bigg\{ 1 + \int_1^\infty p^{1-t}\, dt \bigg\} = \Big(1 + \frac{1}{\log p}\Big) C \sqrt{\frac{\ell \log p}{n}},
\end{aligned}$$

as required. □

Lemma 2.9. *Let $\epsilon \in (0, 1/2)$, let $\ell \in \{1, \dots, p\}$ and let $A \in \mathbb{R}^{p \times p}$ be a symmetric matrix. Then there exists $\mathcal{N}_\epsilon \subseteq \mathbb{S}^{p-1}(\ell)$ with cardinality at most $\binom{p}{\ell} \pi \ell^{1/2} (1 - \epsilon^2/16)^{-(\ell-1)/2} (2/\epsilon)^{\ell-1}$ such that*

$$\sup_{u \in \mathbb{S}^{p-1}(\ell)} |u^\top A u| \le (1 - 2\epsilon)^{-1} \max_{u \in \mathcal{N}_\epsilon} |u^\top A u|.$$

Proof. Let $\mathcal{I}_\ell := \big\{ I \subseteq \{1, \dots, p\} : |I| = \ell \big\}$, and for $I \in \mathcal{I}_\ell$, let $B_I := \{u \in \mathbb{S}^{p-1}(\ell) : u_{I^c} = 0\}$. Thus

$$\mathbb{S}^{p-1}(\ell) = \bigcup_{I \in \mathcal{I}_\ell} B_I.$$

For each $I \in \mathcal{I}_\ell$, by Lemma 10 of Kim and Samworth (2016), there exists $\mathcal{N}_{I,\epsilon} \subseteq B_I$ such that $|\mathcal{N}_{I,\epsilon}| \le \pi \ell^{1/2} (1 - \epsilon^2/16)^{-(\ell-1)/2} (2/\epsilon)^{\ell-1}$ and such that for any $x \in B_I$, there exists $x' \in \mathcal{N}_{I,\epsilon}$ with $\|x - x'\| \le \epsilon$. Let $u_I \in \mathrm{argmax}_{u \in B_I} |u^\top A u|$ and find $v_I \in \mathcal{N}_{I,\epsilon}$ such that $\|u_I - v_I\| \le \epsilon$. Then

$$\begin{aligned}
|u_I^\top A u_I| &\le |v_I^\top A v_I| + |(u_I - v_I)^\top A v_I| + |u_I^\top A (u_I - v_I)| \\
&\le \max_{u \in \mathcal{N}_{I,\epsilon}} |u^\top A u| + 2\epsilon |u_I^\top A u_I|.
\end{aligned}$$

Writing $\mathcal{N}_\epsilon := \cup_{I \in \mathcal{I}_\ell} \mathcal{N}_{I,\epsilon}$, we note that $\mathcal{N}_\epsilon$ has cardinality smaller than or equal to $\binom{p}{\ell} \pi \ell^{1/2} (1 - \epsilon^2/16)^{-(\ell-1)/2} (2/\epsilon)^{\ell-1}$ and that

$$\begin{aligned}
\sup_{u \in \mathbb{S}^{p-1}(\ell)} |u^\top A u| = \max_{I \in \mathcal{I}_\ell} \sup_{u \in B_I} |u^\top A u| &\le (1 - 2\epsilon)^{-1} \max_{I \in \mathcal{I}_\ell} \max_{u \in \mathcal{N}_{I,\epsilon}} |u^\top A u| \\
&= (1 - 2\epsilon)^{-1} \max_{u \in \mathcal{N}_\epsilon} |u^\top A u|,
\end{aligned}$$

as required. □

Lemma 2.10 (Variant of the Gilbert–Varshamov Lemma)**.** *Let $\alpha, \beta \in (0,1)$ and $k, p \in \mathbb{N}$ be such that $k \leq \alpha\beta p$. Writing $\mathcal{S} := \big\{x = (x_1, \ldots, x_p)^\top \in \{0,1\}^p : \sum_{j=1}^p x_j = k\big\}$, there exists a subset $\mathcal{S}_0$ of $\mathcal{S}$ such that for all distinct $x = (x_1, \ldots, x_p)^\top, y = (y_1, \ldots, y_p)^\top \in \mathcal{S}_0$, we have $\sum_{j=1}^p \mathbb{1}_{\{x_j \neq y_j\}} \geq 2(1-\alpha)k$ and such that $\log|\mathcal{S}_0| \geq \rho k \log(p/k)$, where $\rho := \frac{\alpha}{-\log(\alpha\beta)}(-\log\beta + \beta - 1)$.*

Proof. See Massart (2007, Lemma 4.10). □

Let P and Q be two probability measures on a measurable space $(\mathcal{X}, \mathcal{B})$. Recall that if P is absolutely continuous with respect to Q, then the Kullback–Leibler divergence between P and Q is $D(P\|Q) := \int_{\mathcal{X}} \log(dP/dQ)\, dP$, where dP/dQ denotes the Radon–Nikodym derivative of P with respect to Q. If P is not absolutely continuous with respect to Q, we set $D(P\|Q) := \infty$.

Lemma 2.11 (Generalised Fano's Lemma)**.** *Let $P_1, \ldots, P_M$ be probability distributions on a measurable space $(\mathcal{X}, \mathcal{B})$, and assume that $D(P_i\|P_j) \leq \beta$ for all $i \neq j$. Then any measurable function $\hat{\psi} : \mathcal{X} \to \{1, \ldots, M\}$ satisfies*

$$\max_{1 \leq i \leq M} P_i(\hat{\psi} \neq i) \geq 1 - \frac{\beta + \log 2}{\log M}.$$

Proof. See Yu (1997, Lemma 3). □

Lemma 2.12. *Suppose that $P \in \mathcal{P}$ and that $X_1, \ldots, X_n \overset{iid}{\sim} P$. Let $\Sigma := \int_{\mathbb{R}^p} xx^\top\, dP(x)$ and $\hat{\Sigma} := n^{-1}\sum_{i=1}^n X_i X_i^\top$. If $V(u) := \mathbb{E}\{(u^\top X_1)^2\}$ and $\hat{V}(u) := n^{-1}\sum_{i=1}^n (u^\top X_i)^2$ for $u \in \mathbb{S}^{p-1}(2)$, then*

$$\|\hat{\Sigma} - \Sigma\|_\infty \leq 2 \sup_{u \in \mathbb{S}^{p-1}(2)} \big|\hat{V}(u) - V(u)\big|.$$

Proof. Let e_r denote the rth standard basis vector in $\mathbb{R}^p$ and write $X_i = (X_{i,1}, \ldots, X_{i,p})^\top$. Then

$$\begin{aligned}
\|\hat{\Sigma} - \Sigma\|_\infty &= \max_{r,s \in \{1,\ldots,p\}} \Big|\frac{1}{n}\sum_{i=1}^n (X_{i,r}X_{i,s}) - \mathbb{E}(X_{1,r}X_{1,s})\Big| \\
&\leq \max_{r,s \in \{1,\ldots,p\}} \Big|\frac{1}{n}\sum_{i=1}^n \Big\{\Big(\frac{1}{2}e_r + \frac{1}{2}e_s\Big)^\top X_i\Big\}^2 - \mathbb{E}\Big[\Big\{\Big(\frac{1}{2}e_r + \frac{1}{2}e_s\Big)^\top X_1\Big\}^2\Big]\Big| \\
&\quad + \max_{r,s \in \{1,\ldots,p\}} \Big|\frac{1}{n}\sum_{i=1}^n \Big\{\Big(\frac{1}{2}e_r - \frac{1}{2}e_s\Big)^\top X_i\Big\}^2 - \mathbb{E}\Big[\Big\{\Big(\frac{1}{2}e_r - \frac{1}{2}e_s\Big)^\top X_1\Big\}^2\Big]\Big| \\
&\leq 2 \sup_{u \in \mathbb{S}^{p-1}(2)} \big|\hat{V}(u) - V(u)\big|,
\end{aligned}$$

as required. □

Recall the definition of the Graph Vector distribution $\mathrm{GV}_p^g(\pi_0)$ from the proof of Theorem 2.6.

Lemma 2.13. *Let $g = (g_1, \ldots, g_p)^\top \in \{0,1\}^p$, and let $Y_1, \ldots, Y_n$ be independent random vectors, each distributed as $\mathrm{GV}_p^g(\pi_0)$ for some $\pi_0 \in (0, 1/2]$. For any $u \in \mathbb{S}^{p-1}(\ell)$, let $V(u) := \mathbb{E}\{(u^\top Y_1)^2\}$ and $\hat{V}(u) := n^{-1}\sum_{i=1}^n (u^\top Y_i)^2$. Then for every $1 \leq \ell \leq 2/\pi_0$, every $n \in \mathbb{N}$ and every $\delta > 0$,*

$$\mathbb{P}\bigg[\sup_{u \in \mathbb{S}^{p-1}(\ell)} |\hat{V}(u) - V(u)| \geq 750 \max\bigg\{\sqrt{\frac{\ell \log(p/\delta)}{n}}, \frac{\ell \log(p/\delta)}{n}\bigg\}\bigg] \leq \delta.$$

In other words, $\mathrm{GV}_p^g(\pi_0) \in \mathrm{RCC}_p(\ell, 750)$ for all $\pi_0 \in (0, 1/2]$ and $\ell \leq 2/\pi_0$.

Proof. We can write

$$Y_i = \xi_i\big\{(1-\epsilon_i)R_i + \epsilon_i(g + \tilde{R}_i)\big\},$$

where ξ_i, ϵ_i and R_i are independent, where ξ_i is a Rademacher random variable, where $\epsilon_i \sim \mathrm{Bern}(\pi_0)$, where $R_i = (r_{i1}, \ldots, r_{ip})^\top$ has independent Rademacher coordinates, and where $\tilde{R}_i = (\tilde{r}_{i1}, \ldots, \tilde{r}_{ip})^\top$ with $\tilde{r}_{ij} := (1-g_j)r_{ij}$. Thus, for any $u \in \mathbb{S}^{p-1}(\ell)$, we have

$$(u^\top Y_i)^2 = (1-\epsilon_i)(u^\top R_i)^2 + \epsilon_i(u^\top g)^2 + \epsilon_i(u^\top \tilde{R}_i)^2 + 2\epsilon_i(u^\top \tilde{R}_i)(u^\top g).$$

Hence, writing $S := \{j : g_j = 1\}$, and using the definition of $\hat{V}(u)$, we have

$$\begin{aligned}|\hat{V}(u) - V(u)| \leq \bigg|\frac{1}{n}\sum_{i=1}^n (1-\epsilon_i)\big\{(u^\top R_i)^2 - 1\big\}\bigg| + \frac{1 + (u^\top g)^2 + \|u_{S^c}\|_2^2}{n}\bigg|\sum_{i=1}^n(\epsilon_i - \pi_0)\bigg| \\ + \bigg|\frac{1}{n}\sum_{i=1}^n \epsilon_i\big\{(u^\top \tilde{R}_i)^2 - \|u_{S^c}\|_2^2\big\}\bigg| + \bigg|\frac{2u^\top g}{n}\sum_{i=1}^n \epsilon_i(u^\top \tilde{R}_i)\bigg|. \qquad (2.30)\end{aligned}$$

We now control the four terms on the right-hand side of (2.30) separately. For the first term, note that the distribution of R_i is subgaussian with parameter 1. Writing $N_{\boldsymbol{\epsilon}} := \sum_{i=1}^n \epsilon_i$, it follows by the same argument as in the proof of Proposition 2.1(i) that for any $s > 0$,

$$\begin{aligned}&\mathbb{P}\bigg(\sup_{u\in\mathbb{S}^{p-1}(\ell)}\bigg|\frac{1}{n}\sum_{i=1}^n(1-\epsilon_i)\big\{(u^\top R_i)^2 - 1\big\}\bigg| \geq 2s\bigg) \\ &\quad= \mathbb{E}\bigg\{\mathbb{P}\bigg(\sup_{u\in\mathbb{S}^{p-1}(\ell)}\bigg|\frac{1}{n-N_{\boldsymbol{\epsilon}}}\sum_{i:\epsilon_i=0}\big\{(u^\top R_i)^2 - 1\big\}\bigg| \geq \frac{2ns}{n-N_{\boldsymbol{\epsilon}}}\;\bigg|\; N_{\boldsymbol{\epsilon}}\bigg)\bigg\} \\ &\quad\leq e^9 p^\ell \mathbb{E}\bigg[\exp\bigg\{-\frac{n(\frac{ns}{n-N_{\boldsymbol{\epsilon}}})^2}{4(\frac{ns}{n-N_{\boldsymbol{\epsilon}}}) + 32}\bigg\}\bigg] \leq e^9 p^\ell \exp\bigg(-\frac{ns^2}{4s+32}\bigg).\end{aligned}$$

We deduce that for any $\delta > 0$,

$$\mathbb{P}\bigg(\sup_{u\in\mathbb{S}^{p-1}(\ell)}\bigg|\frac{1}{n}\sum_{i=1}^n(1-\epsilon_i)\big\{(u^\top R_i)^2 - 1\big\}\bigg| \geq 16\max\bigg\{\sqrt{\frac{\ell\log(p/\delta)}{n}}, \frac{\ell\log(p/\delta)}{n}\bigg\}\bigg) \leq e^9\delta. \qquad (2.31)$$

For the second term on the right-hand side of (2.30), note first that for any $u \in \mathbb{S}^{p-1}(\ell)$, we have by Cauchy–Schwarz that

$$(u^\top g)^2 \leq \|u_S\|_0\|u_S\|_2^2 \leq \|u_S\|_0 \leq \ell.$$

We deduce using Bernstein's inequality for Binomial random variables (e.g. Shorack and Wellner, 1986, p. 855) that for any $s > 0$,

$$\begin{aligned}&\mathbb{P}\bigg\{\sup_{u\in\mathbb{S}^{p-1}(\ell)}\frac{1 + (u^\top g)^2 + \|u_{S^c}\|_2^2}{n}\bigg|\sum_{i=1}^n(\epsilon_i - \pi_0)\bigg| \geq s\bigg\} \leq \mathbb{P}\bigg\{\frac{1}{n}\bigg|\sum_{i=1}^n(\epsilon_i - \pi_0)\bigg| \geq \frac{s}{3\ell}\bigg\} \\ &\quad\leq 2\exp\bigg(-\frac{ns^2}{18\ell^2\pi_0 + 2s\ell}\bigg) \leq 2\max\bigg\{\exp\bigg(-\frac{ns^2}{(19+\sqrt{37})\ell^2\pi_0}\bigg), \exp\bigg(-\frac{ns}{(1+\sqrt{37})\ell}\bigg)\bigg\}.\end{aligned}$$

By assumption, $\ell\pi_0 \leq 2$. Hence, for any $\delta > 0$,

$$\mathbb{P}\Big\{\sup_{u\in\mathbb{S}^{p-1}(\ell)} \frac{1+(u^\top g)^2+\|u_{S^c}\|_2^2}{n}\Big|\sum_{i=1}^n(\epsilon_i-\pi_0)\Big| \geq (1+\sqrt{37})\max\Big(\sqrt{\frac{\ell\log(1/\delta)}{n}}, \frac{\ell\log(1/\delta)}{n}\Big)\Big\} \leq 2\delta. \tag{2.32}$$

The third term on the right-hand side of (2.30) can be handled in a very similar way to the first. We find that for every $\delta > 0$,

$$\mathbb{P}\Big(\sup_{u\in\mathbb{S}^{p-1}(\ell)}\Big|\frac{1}{n}\sum_{i=1}^n \epsilon_i\big\{(u^\top\tilde{R}_i)^2 - \|u_{S^c}\|_2^2\big\}\Big| \geq 16\max\Big\{\sqrt{\frac{\ell\log(p/\delta)}{n}}, \frac{\ell\log(p/\delta)}{n}\Big\}\Big) \leq e^9\delta. \tag{2.33}$$

For the final term, by definition of $\tilde{R}_i$, we have for any $u \in \mathbb{S}^{p-1}(\ell)$ that

$$\Big|\frac{2u^\top g}{n}\sum_{i=1}^n \epsilon_i(u^\top\tilde{R}_i)\Big| \leq \frac{2\ell^{1/2}}{n}\Big|\sum_{j:g_j=0} u_j \sum_{i:\epsilon_i=1} r_{ij}\Big| \leq \frac{2\ell}{n}\max_{j:g_j=0}\Big|\sum_{i:\epsilon_i=1} r_{ij}\Big|.$$

Hence by Hoeffding's inequality, for any $s > 0$,

$$\begin{aligned}
\mathbb{P}\Big\{\sup_{u\in\mathbb{S}^{p-1}(\ell)}\Big|\frac{2u^\top g}{n}\sum_{i=1}^n\epsilon_i(u^\top\tilde{R}_i)\Big| \geq s\Big\} &\leq \mathbb{E}\Big\{\mathbb{P}\Big(\max_{1\leq j\leq p}\Big|\sum_{i:\epsilon_i=1} r_{ij}\Big| \geq \frac{ns}{2\ell}\;\Big|\; N_{\boldsymbol{\epsilon}}\Big)\Big\} \\
&\leq 2p\mathbb{E}\Big\{\exp\Big(-\frac{n^2s^2}{8\ell^2 N_{\boldsymbol{\epsilon}}}\Big)\Big\} \leq 2p\inf_{t>0}\Big\{\exp\Big(-\frac{n^2s^2}{8\ell^2 t}\Big) + \mathbb{P}(N_{\boldsymbol{\epsilon}} > t)\Big\} \\
&\leq 2p\inf_{t>0}\Big\{\exp\Big(-\frac{n^2s^2}{8\ell^2 t}\Big) + \exp\Big(-t\log\frac{t}{n\pi_0} + t - n\pi_0\Big)\Big\},
\end{aligned}$$

where the final line follows by Bennett's inequality (e.g. Shorack and Wellner, 1986, p. 440). Choosing $t = \max\big(e^2 n\pi_0, \frac{ns}{2^{3/2}\ell}\big)$, we find

$$\begin{aligned}
&\mathbb{P}\Big\{\sup_{u\in\mathbb{S}^{p-1}(\ell)}\Big|\frac{2u^\top g}{n}\sum_{i=1}^n\epsilon_i(u^\top\tilde{R}_i)\Big| \geq s\Big\} \\
&\qquad\leq 2p\max\Big\{\exp\Big(-\frac{ns^2}{8e^2\ell^2\pi_0}\Big) + \exp\Big(-\frac{ns}{2^{3/2}\ell}\Big), 2\exp\Big(-\frac{ns}{2^{3/2}\ell}\Big)\Big\} \\
&\qquad\leq 4p\max\Big\{\exp\Big(-\frac{ns^2}{16e^2\ell}\Big), \exp\Big(-\frac{ns}{2^{3/2}\ell}\Big)\Big\}.
\end{aligned}$$

We deduce that for any $\delta > 0$,

$$\mathbb{P}\Big[\sup_{u\in\mathbb{S}^{p-1}(\ell)}\Big|\frac{2u^\top g}{n}\sum_{i=1}^n\epsilon_i(u^\top\tilde{R}_i)\Big| \geq 4e\max\Big\{\sqrt{\frac{\ell\log(p/\delta)}{n}}, \frac{\ell\log(p/\delta)}{n}\Big\}\Big] \leq 4\delta. \tag{2.34}$$

We conclude from (2.30), (2.31), (2.32), (2.33) and (2.34) that for any $\delta > 0$,

$$\mathbb{P}\Big[\sup_{u\in\mathbb{S}^{p-1}(\ell)}|\hat{V}(u) - V(u)| \geq 750\max\Big\{\sqrt{\frac{\ell\log(p/\delta)}{n}}, \frac{\ell\log(p/\delta)}{n}\Big\}\Big] \leq \delta,$$

as required. □

Lemma 2.14. *Let $v = (v_1, \ldots, v_p)^\top \in \mathbb{S}^{p-1}(k)$ and let $\hat{v} = (\hat{v}_1, \ldots, \hat{v}_p)^\top \in \mathbb{R}^p$ be a unit vector.*

Let $S := \{j \in \{1, \dots, p\} : v_j \neq 0\}$. *Then for any* $\hat{S} \in \operatorname{argmax}_{1 \leq j_1 < \dots < j_k \leq p} \sum_{r=1}^{k} |\hat{v}_{j_r}|$, *we have*

$$L(\hat{v}, v)^2 \geq \frac{1}{2} \sum_{j \in S \setminus \hat{S}} v_j^2.$$

Proof. By the Cauchy–Schwarz inequality, and then by definition of $\hat{S}$,

$$\begin{aligned} 1 - L(\hat{v}, v)^2 &= \Big(\sum_{j \in S \setminus \hat{S}} \hat{v}_j v_j + \sum_{j \in S \cap \hat{S}} \hat{v}_j v_j \Big)^2 \leq \Big(2 \sum_{j \in S \setminus \hat{S}} \hat{v}_j^2 + \sum_{j \in S \cap \hat{S}} \hat{v}_j^2 \Big) \Big(\frac{1}{2} \sum_{j \in S \setminus \hat{S}} v_j^2 + \sum_{j \in S \cap \hat{S}} v_j^2 \Big) \\ &\leq \Big(\sum_{j \in \hat{S} \setminus S} \hat{v}_j^2 + \sum_{j \in S \setminus \hat{S}} \hat{v}_j^2 + \sum_{j \in S \cap \hat{S}} \hat{v}_j^2 \Big) \Big(1 - \frac{1}{2} \sum_{j \in S \setminus \hat{S}} v_j^2 \Big) \leq 1 - \frac{1}{2} \sum_{j \in S \setminus \hat{S}} v_j^2, \end{aligned}$$

as required. □

Lemma 2.15. *Let* $A \in \mathbb{R}^{d \times d}$ *be a symmetric matrix. Let* $A^{(r)}$ *be the principal submatrix of* A *obtained by deleting the* r*th row and* r*th column of* A. *If* A *has a unique (up to sign) leading eigenvector* v, *then*

$$\lambda_2(A) \leq \lambda_1(A^{(r)}) \leq \lambda_1(A) - v_{1,r}^2 (\lambda_1(A) - \lambda_2(A))$$

Proof. The first inequality in the lemma is implied by Cauchy's Interlacing Theorem (see, e.g. Horn and Johnson (2012, Theorem 4.3.17)). It remains to show the second inequality. Let $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_d$ be eigenvalues of A (counting multiplicities), and $v_1, \dots, v_d$ be unit-length eigenvectors of A such that $Av_i = \lambda_i v_i$ and $v_i^\top v_j = 0$ for all $i \neq j$. We have

$$\begin{aligned} \lambda_1(A^{(r)}) &= \max_{\substack{\|u\|_2 = 1 \\ u_r = 0}} u^\top A u = \max_{\substack{\|u\|_2 = 1 \\ u_r = 0}} u^\top \Big(\sum_{i=1}^{d} \lambda_i v_i v_i^\top \Big) u \\ &\leq \max_{\substack{\|u\|_2 = 1 \\ u_r = 0}} \Big\{ (\lambda_1 - \lambda_2) u^\top v_1 v_1^\top u + \lambda_2 u^\top \Big(\sum_{i=1}^{d} v_i v_i^\top \Big) u \Big\} \\ &\leq \max_{\substack{\|u\|_2 = 1 \\ u_r = 0}} (\lambda_1 - \lambda_2) |u^\top v_1|^2 + \lambda_2 \leq (\lambda_1 - \lambda_2)(1 - v_{1,r}^2) + \lambda_2 = \lambda_1 - v_{1,r}^2 (\lambda_1 - \lambda_2), \end{aligned}$$

where we used Cauchy–Schwarz inequality in the penultimate line. □

Recall the definition of the total variation distance d_{TV} given in the proof of Theorem 2.6.

Lemma 2.16. *Let* X *and* Y *be random elements taking values in a measurable space* $(F, \mathcal{F})$, *and let* $(G, \mathcal{G})$ *be another measurable space.*

(a) If $\phi : F \to G$ *is measurable, then* $d_{\mathrm{TV}}\big(\mathcal{L}(\phi(X)), \mathcal{L}(\phi(Y))\big) \leq d_{\mathrm{TV}}\big(\mathcal{L}(X), \mathcal{L}(Y)\big)$.

(b) Let Z *be a random element taking values in* $(G, \mathcal{G})$, *and suppose that* Z *is independent of* (X, Y). *Then* $d_{\mathrm{TV}}\big(\mathcal{L}(X, Z), \mathcal{L}(Y, Z)\big) = d_{\mathrm{TV}}\big(\mathcal{L}(X), \mathcal{L}(Y)\big)$.

Proof. (a) For any $A \in \mathcal{G}$, we have

$$|\mathbb{P}\{\phi(X) \in A\} - \mathbb{P}\{\phi(Y) \in A\}| = |\mathbb{P}\{X \in \phi^{-1}(A)\} - \mathbb{P}\{Y \in \phi^{-1}(A)\}| \leq d_{\mathrm{TV}}\big(\mathcal{L}(X), \mathcal{L}(Y)\big).$$

Since $A \in \mathcal{G}$ was arbitrary, the result follows.

(b) Define $\phi : F \times G \to F$ by $\phi(w, z) := w$. Then ϕ is measurable. By part (a), we have

$$d_{\mathrm{TV}}\big(\mathcal{L}(X), \mathcal{L}(Y)\big) = d_{\mathrm{TV}}\big(\mathcal{L}(\phi(X, Z)), \mathcal{L}(\phi(Y, Z))\big) \leq d_{\mathrm{TV}}\big(\mathcal{L}(X, Z), \mathcal{L}(Y, Z)\big).$$

For the other inequality, let $\mathcal{A}$ denote the set of subsets A of $\mathcal{F} \otimes \mathcal{G}$ with the property that given $\epsilon > 0$, there exist sets $B_{1,F}, \ldots, B_{n,F} \in \mathcal{F}$ and disjoint sets $B_{1,G}, \ldots, B_{n,G} \in \mathcal{G}$ such that, writing $B := \cup_{i=1}^{n}(B_{i,F} \times B_{i,G})$, we have $\mathbb{P}\big((X, Z) \in A \triangle B\big) < \epsilon$ and $\mathbb{P}\big((Y, Z) \in A \triangle B\big) < \epsilon$. Here, the binary operator $\triangle$ denotes the symmetric difference of two sets, so that $A \triangle B := (A \cap B^c) \cup (A^c \cap B)$. Note that $\mathcal{F} \times \mathcal{G} \subseteq \mathcal{A}$. Now suppose $A \in \mathcal{A}$ so that, given $\epsilon > 0$, we can find sets $B_{1,F}, \ldots, B_{n,F} \in \mathcal{F}$ and disjoint sets $B_{1,G}, \ldots, B_{n,G} \in \mathcal{G}$ with the properties above. Observe that we can write

$$B^c = \bigcup_{I \subseteq \{1,\ldots,n\}} \Big(\bigcap_{i \in I} B_{i,F}^c \times \bigcap_{i \in I} B_{i,G} \cap \bigcap_{i \in I^c} B_{i,G}^c \Big).$$

For each $I \subseteq \{1, \ldots, n\}$, the sets $\cap_{i \in I} B_{i,F}^c$ belong to $\mathcal{F}$, and $\big\{\cap_{i \in I} B_{i,G} \cap \cap_{i \in I^c} B_{i,G}^c : I \subseteq \{1, \ldots, n\}\big\}$ is a family of disjoint sets in $\mathcal{G}$. Moreover,

$$\mathbb{P}\big((X, Z) \in A^c \triangle B^c\big) = \mathbb{P}\big((X, Z) \in A \triangle B\big) < \epsilon,$$

and similarly $\mathbb{P}\big((Y, Z) \in A^c \triangle B^c\big) < \epsilon$. We deduce that $A^c \in \mathcal{A}$. Finally, if (A_n) is a disjoint sequence in $\mathcal{A}$, then let $A := \cup_{n=1}^{\infty} A_n$, and given $\epsilon > 0$, find $m \in \mathbb{N}$ such that $\mathbb{P}\big((X, Z) \in A \setminus \cup_{i=1}^{m} A_i\big) < \epsilon/2$ and $\mathbb{P}\big((Y, Z) \in A \setminus \cup_{i=1}^{m} A_i\big) < \epsilon/2$. Now, for each $i = 1, \ldots, m$, find sets $B_{i1,F}, \ldots, B_{in_i,F} \in \mathcal{F}$ and disjoint sets $B_{i1,G}, \ldots, B_{in_i,G} \in \mathcal{G}$ such that, writing $B_i := \cup_{j=1}^{n_i}(B_{ij,F} \times B_{ij,G})$, we have $\mathbb{P}\big((X, Z) \in A_i \triangle B_i\big) < \epsilon/(2m)$ and $\mathbb{P}\big((Y, Z) \in A_i \triangle B_i\big) < \epsilon/(2m)$. It is convenient to relabel the sets $\{(B_{ij,F}, B_{ij,G}) : i = 1, \ldots, m, j = 1, \ldots, n_i\}$ as $\{(C_{1,F}, C_{1,G}), \ldots, (C_{N,F}, C_{N,G})\}$, where $N := \sum_{i=1}^{m} n_i$. This means that we can write

$$\bigcup_{i=1}^{m} B_i = \bigcup_{k=1}^{N} (C_{k,F} \times C_{k,G}) = \bigcup_{K \subseteq \{1,\ldots,N\}, K \neq \emptyset} \Big(\bigcup_{k \in K} C_{k,F} \times \bigcap_{k \in K} C_{k,G} \cap \bigcap_{k \in K^c} C_{k,G}^c \Big).$$

Now, for each non-empty subset K of $\{1, \ldots, N\}$, the set $\cup_{k \in K} C_{k,F}$ belongs to $\mathcal{F}$, and $\big\{\cap_{k \in K} C_{k,G} \cap \cap_{k \in K^c} C_{k,G}^c : K \subseteq \{1, \ldots, N\}, K \neq \emptyset\big\}$ is a family of disjoint sets in $\mathcal{G}$. Moreover,

$$\mathbb{P}\big((X, Z) \in A \triangle \cup_{i=1}^{m} B_i\big) \leq \sum_{i=1}^{m} \mathbb{P}\big((X, Z) \in A_i \triangle B_i\big) + \frac{\epsilon}{2} < \epsilon,$$

and similarly, $\mathbb{P}\big((Y, Z) \in A \triangle \cup_{i=1}^{m} B_i\big) < \epsilon$. We deduce that $A \in \mathcal{A}$, so $\mathcal{A}$ is a σ-algebra containing $\mathcal{F} \times \mathcal{G}$, so $\mathcal{A}$ contains $\mathcal{F} \otimes \mathcal{G}$.

Now suppose that $A \in \mathcal{F} \otimes \mathcal{G}$. By the argument above, given $\epsilon > 0$, there exist sets $B_{1,F}, \ldots, B_{n,F} \in \mathcal{F}$ and disjoint sets $B_{1,G}, \ldots, B_{n,G} \in \mathcal{G}$ such that $\mathbb{P}\big((X, Z) \in A \triangle \cup_{i=1}^{m} (B_{i,F} \times B_{i,G})\big) < \epsilon/2$ and $\mathbb{P}\big((Y, Z) \in A \triangle \cup_{i=1}^{m} (B_{i,F} \times B_{i,G})\big) < \epsilon/2$. It follows that

$$\begin{aligned}\big|\mathbb{P}\big((X, Z) \in A\big) - \mathbb{P}\big((Y, Z) \in A\big)\big| &\leq \sum_{i=1}^{m} \big|\mathbb{P}\big(X \in B_{i,F}, Z \in B_{i,G}\big) - \mathbb{P}\big(Y \in B_{i,F}, Z \in B_{i,G}\big)\big| + \epsilon \\ &= \sum_{i=1}^{m} \mathbb{P}(Z \in B_{i,G}) \big|\mathbb{P}(X \in B_{i,F}) - \mathbb{P}(Y \in B_{i,F})\big| + \epsilon \leq d_{\mathrm{TV}}\big(\mathcal{L}(X), \mathcal{L}(Y)\big) + \epsilon.\end{aligned}$$

Since $A \in \mathcal{A}$ and $\epsilon > 0$ were arbitrary, we conclude that

$$d_{\mathrm{TV}}\big(\mathcal{L}(X,Z),\mathcal{L}(Y,Z)\big) \leq d_{\mathrm{TV}}\big(\mathcal{L}(X),\mathcal{L}(Y)\big),$$

as required. □

2.8 Appendix: A brief introduction to computational complexity theory

The following is intended to give a short introduction to notions in computational complexity theory. A good reference for further information is Arora and Barak (2009), from which much of the following is inspired.

A *computational problem* is the task of generating a desired output based on a given input. Formally, defining $\{0,1\}^* := \cup_{k=1}^{\infty}\{0,1\}^k$ to be the set of all finite strings of zeros and ones, we can view a computational problem as a function $F : \{0,1\}^* \to \mathcal{P}\big(\{0,1\}^*\big)$, where $\mathcal{P}(A)$ denotes the power set of a set A. The interpretation is that $F(s)$ describes the set of acceptable output strings (solutions) for a particular input string s.

Loosely speaking, an *algorithm* is a collection of instructions for performing a task. Despite the widespread use of algorithms in mathematics throughout history, it was not until 1936 that Alonzo Church and Alan Turing formalised the notion by defining notational systems called the λ-calculus and Turing machines respectively (Church, 1936; Turing, 1936). Here we define an algorithm to be a *Turing machine*:

Definition 2.1. A *Turing machine* M is a pair (Q,δ), where

- Q is a finite set of states, among which are two distinguished states q_{start} and q_{halt}.
- δ is a 'transition' function from $Q \times \{0,1,\text{␣}\}$ to $Q \times \{0,1,\text{␣}\} \times \{\mathrm{L},\mathrm{R}\}$.

A Turing Machine can be thought of as having a reading head that can access a tape consisting of a countably infinite number of squares, labelled $0,1,2,\ldots$. When the Turing machine is given an input $s \in \{0,1\}^*$, the tape is initialised with the components of s in its first $|s|$ tape squares (where $|\cdot|$ denotes the length of a string in $\{0,1\}^*$) and with 'blank symbols' ␣ in its remaining squares. The Turing machine starts in the state $q_{\mathrm{start}} \in Q$ with its head on the 0th square and operates according to its transition function δ. When the machine is in state $q \in Q$ with its head over the ith tape square that contains the symbol $a \in \{0,1,\text{␣}\}$, and if $\delta(q,a) = (q',a',\mathrm{L})$, the machine overwrites a with a', updates its state to q', and moves to square $i-1$ (or to square $i+1$ if the third component of the transition function is R instead of L). The Turing machine stops if it reaches state $q_{\mathrm{halt}} \in Q$ and outputs the vector of symbols on the tape before the first blank symbol. If the Turing machine M terminates (in finitely many steps) with input s, we write $M(s)$ for its output.

We say an algorithm (Turing machine) *M solves a computational problem F* if M terminates for every input $s \in \{0,1\}^*$, and $M(s) \in F(s)$. A computational problem is *solvable* if there exists a Turing machine that solves it. It turns out that other notions of an algorithm (including Church's λ-calculus and modern computer programming languages) are equivalent in the sense that the set of solvable problems is the same.

A *polynomial time algorithm* is a Turing machine M for which there exist $a, b > 0$ such that for all input strings $s \in \{0,1\}^*$, M terminates after at most $a|s|^b$ transitions. We say a problem F is *polynomial time solvable*, written $F \in \mathsf{P}$, if there exists a polynomial time algorithm that solves it[5].

A *nondeterministic Turing machine* has the same definition as that for a Turing machine except that the transition function δ becomes a set-valued function $\delta : Q \times \{0,1,\textvisiblespace\} \to \mathcal{P}\big(Q \times \{0,1,\textvisiblespace\} \times \{\mathrm{L},\mathrm{R}\}\big)$. The idea is that, while in state q with its head over symbol a, a nondeterministic Turing machine replicates $|\delta(q,a)|$ copies of itself (and its tape) in the current configuration, each exploring a different possible future configuration in the set $\delta(q,a)$. Each replicate branches to further replicates in the next step. The process continues until one of its replicates reaches the state q_{halt}. At that point, the Turing machine replicate that has halted outputs its tape content and all replicates stop computation. A *nondeterministic polynomial time algorithm* is a nondeterministic Turing machine M_{nd} for which there exist $a, b > 0$ such that for all input strings $s \in \{0,1\}^*$, M_{nd} terminates after at most $a|s|^b$ steps. (We count all replicates of M_{nd} making one parallel transition as one step.) We say a computational problem F is *nondeterministically polynomial time solvable*, written $F \in \mathsf{NP}$, if there exists a nondeterministic polynomial time algorithm that solves it[6].

Clearly $\mathsf{P} \subseteq \mathsf{NP}$, but it is not currently known if these classes are equal. It is widely believed that $\mathsf{P} \neq \mathsf{NP}$, and many computational lower bounds for particular computational problems have been proved conditional under this assumption. Working under this hypothesis, a common strategy is to relate the algorithmic complexity of one computational problem to another. We say a computational problem F is *polynomial time reducible* to another problem G, written as $F \leq_{\mathsf{P}} G$, if there exist polynomial time algorithms M_{in} and M_{out} such that $M_{\text{out}} \circ G \circ M_{\text{in}}(s) \subseteq F(s)$. In other words, $F \leq_{\mathsf{P}} G$ if we can convert an input of F to an input of G through M_{in}, and translate every solution of G back to a solution for F through M_{out}.

Definition 2.2. A computational problem G is NP-*hard* if $F \leq_{\mathsf{P}} G$ for all $F \in \mathsf{NP}$. It is NP-*complete* if it is in NP and is NP-hard.

Karp (1972) showed that a large number of natural computational problems are NP-complete, including the Clique problem mentioned in Section 2.4. The Turing machines and nondeterministic Turing machines introduced above are both non-random. In some situations (e.g. statistical problems), it is useful to consider random procedures:

Definition 2.3. A *probabilistic Turing machine* M_{pr} is a triple (Q, δ, X), where

- Q is a finite set of states, among which are two distinguished states q_{start} and q_{halt}.
- δ is a transition function from $Q \times \{0,1,\textvisiblespace\} \times \{0,1\}$ to $Q \times \{0,1,\textvisiblespace\} \times \{\mathrm{L},\mathrm{R}\}$.
- $X = (X_1, X_2, \ldots)$ is an infinite sequence of independent Bern(1/2) random variables.

In its tth step, if a probabilistic Turing machine M_{pr} is in state q with its reading head over symbol a, and $\delta(q, a, X_t) = (q', a', L)$, then M_{pr} overwrites a with a', updates its state to q' and moves its reading head to the left (or to the right if $\delta(q, a, X_t) = (q', a', R)$). A *randomised polynomial time algorithm* is a probabilistic Turing machine M_{pr} for which there exist $a, b > 0$ such

[5] In fact, some authors write FP (short for 'Functional Polynomial Time') for the class we have denoted as P here. The notation P is then reserved for the subset of computational problems consisting of so-called *decision problems* F, where $F(s) \in \big\{\{0\},\{1\}\big\}$ for all $s \in \{0,1\}^*$.

[6] Again, some authors write FNP for the class we have denoted as NP here.

that for any $s \in \{0,1\}^*$, M_{pr} terminates in at most $a|s|^b$ steps. We say a computational problem F is *solvable in randomised polynomial time*, written as $F \in \mathsf{BPP}$, if, given $\epsilon > 0$, there exists a randomised polynomial time algorithm $M_{\text{pr},\epsilon}$ such that $\mathbb{P}\big(M_{\text{pr},\epsilon}(s) \in F(s)\big) \geq 1 - \epsilon$.

In the above discussion, the classes P, NP, BPP are all defined through worst-case performance of an algorithm, since we require the time bound to hold for every input string s. However, in many statistical applications, the input string s is drawn from some distribution $\mathcal{D}$ on $\{0,1\}^*$, and it is the average performance of the algorithm, rather than the worst case scenario, that is of more interest. We say such a random problem is solvable in randomised polynomial time if, given $\epsilon > 0$, there exists a randomised polynomial time algorithm $M_{\text{pr},\epsilon}$ such that, when $s \sim \mathcal{D}$, independent of X, we have $\mathbb{P}\big(M_{\text{pr}}(s) \in F(s)\big) \geq 1 - \epsilon$. Note that the probability here is taken over both the randomness in s and the randomness in X. Similar to the non-random cases, we can talk about randomised polynomial time reduction. If M_F is a randomised polynomial time algorithm for a computational problem F, then $M_{\text{out}} \circ M_F \circ M_{\text{in}}$ is a potential randomised polynomial time algorithm for another problem G for suitably constructed randomised polynomial time algorithms M_{in} and M_{out}. One such construction is the key to the proof of Theorem 2.6.

Chapter 3

Average-case hardness of restricted isometry certification

3.1 Introduction

In many areas of data science, high-dimensional signals contain rich structure. It is of great interest to leverage this structure to improve our ability to describe characteristics of the signal and to make future predictions. Sparsity is a structure of wide applicability (see, e.g. Mallat, 1999; Rauhut and Foucart, 2013; Eldar and Kutyniok, 2012), with a broad literature dedicated to its study in various scientific fields.

The sparse linear model takes the form $y = X\beta + \varepsilon$, where $y \in \mathbb{R}^n$ is a vector of observations, $X \in \mathbb{R}^{n\times p}$ is a design matrix, $\varepsilon \in \mathbb{R}^n$ is noise and the vector $\beta \in \mathbb{R}^p$ is assumed to have a small number k of non-zero entries. Estimating β or the mean response, $X\beta$, are among the most widely studied problems in signal processing, as well as in statistical learning. In high-dimensional problems, one would wish to recover β with as few observations as possible. For an incoherent design matrix, it is known that an order of k^2 observations suffice (Donoho and Elad, 2003; Donoho, Elad and Temlyakov, 2006). However, this appears to require a number of observations far exceeding the information content of β, which has only k variables, albeit with unknown locations.

This dependence in k can be greatly improved by using design matrices that are almost isometries on some low dimensional subspaces, i.e., matrices that satisfy the *restricted isometry property* with parameters k and θ, or $\mathrm{RIP}_{n,p}(k,\theta)$ (see Definition 3.1). It is a highly robust property, and in fact implies that many different polynomial time algorithms, such as greedy methods (Blumensath and Davies, 2009; Dai and Milenkovic, 2009; Needell and Tropp, 2009) and convex optimisation (Candès and Tao, 2005; Candès, Romberg and Tao, 2006b; Candès and Tao, 2006; Candès, 2008), are stable in recovering β. Random matrices are known to satisfy the RIP when the number n of observation is of higher order than $k\log(p)/\theta^2$. These results were developed in the field of compressed sensing (Candès, Romberg and Tao, 2006a; Candès and Tao, 2006; Rauhut and Foucart, 2013; Donoho, 2006; Eldar and Kutyniok, 2012), where the use of randomness is pivotal for near-optimal results. While the assumption of randomness allows great theoretical leaps, it leaves open questions for practitioners.

Scientists working on data closely following this model cannot always choose their design matrix

X, or at least choose one that is completely random. Moreover, it is in general practically impossible to check that a given matrix satisfies these desired properties, as RIP certification is NP-hard (Bandeira et al., 2012). Having access to a function, or a statistic, of X that could be easily computed, which determines how well β may be estimated, would therefore be of great help.

The search for such statistics has been of great importance for over a decade now, and several have been proposed (d'Aspremont, Bach and El Ghaoui, 2008; Lee and Bresler, 2008; d'Aspremont and El Ghaoui, 2011; Juditsky and Nemirovski, 2011). Perhaps the simplest and most popular is the incoherence parameter, which measures the maximum inner product between distinct, normalised, columns of X. However, all of these are known to necessarily fail to guarantee good recovery when $p \geq 2n$ unless n is of order k^2 (d'Aspremont and El Ghaoui, 2011). Given a specific problem instance, the strong recovery guarantees of compressed sensing cannot be verified based on these statistics.

In this chapter, we study the problem of *average-case certification* of the Restricted Isometry Property (RIP). A certifier takes as input a design matrix X, always outputs 'false' when X does not satisfy the property, and outputs 'true' for a large proportion of matrices (see Definition 3.4). Indeed, worst-case hardness does not preclude a problem from being solvable for most instances. The link between restricted isometry and incoherence implies that polynomial time certifiers exists in a regime where n is of order $k^2 \log(p)/\theta^2$. It is natural to ask whether the RIP can be certified for sample size $n \gg k \log(p)/\theta^2$, where most matrices (with respect to, say, the Gaussian measure) are RIP. If it does, it would also provide a Las Vegas algorithm to construct RIP design matrices of optimal sizes. This should be compared with the currently existing limitations for the deterministic construction of RIP matrices.

Our main result is that certification in this average sense is computationally hard even in a near-optimal parameter regime, assuming a new, weaker assumption on detecting dense subgraphs, related to the Planted Clique hypothesis.

Theorem (Informal). *There is no computationally efficient, average-case certifier for the class $RIP_{n,p}(k, \theta)$ uniformly over an asymptotic regime where $n \ll k^{1+\alpha}/\theta^2$, for any $\alpha < 1$.*

This suggests that even in the average case, RIP certification requires almost $k^2 \log(p)/\theta^2$ observations. This contrasts highly with the fact that a random matrix satisfies RIP with high probability when n exceeds about $k \log(p)/\theta^2$. Thus, there appears to be a large gap between what a practitioner may be able to certify given a specific problem instance, and what holds for a random matrix. On the other hand, if a certifier is found which fills this gap, the result would not only have huge practical implications in compressed sensing and statistical learning, but would also disprove a long-standing conjecture from computational complexity theory.

Our result shares many characteristics with a hypothesis by Feige (2002) on the hardness of refuting random satisfiability formulas. Indeed, our statement is also about the hardness of verifying that a property holds for a particular instance (RIP for design matrices, instead of unsatisfiability for boolean formulas). It concerns a regime where such a property should hold with high probability (n of order $k^{1+\alpha}/\theta^2$, linear regime for satisfiability), cautiously allowing only one type of errors, false negatives, for a problem that is hard in the worst case. In these two examples, such certifiers exist in a sub-optimal regime. Our problem is conceptually different from results regarding the worst-case hardness of certifying this property (see, e.g. Bandeira et al., 2012; Koiran and Zouzias, 2012; Tillmann and Pfetsch, 2014). It is closer to another line of work concerned with computational lower bounds for statistical learning problems based on average-case

assumptions. The planted clique assumption has been used to prove computational hardness results for statistical problems such as estimation and testing of sparse principal components (Berthet and Rigollet (2013a,b), see also Chapter 2), testing and localisation of submatrix signals (Ma and Wu, 2015; Chen and Xu, 2016), community detection (Hajek, Wu and Xu, 2015) and sparse canonical correlation analysis (Gao, Ma and Zhou, 2014). The intractability of noisy parity recovery problem (Blum, Kalai and Wasserman, 2003) has also been used recently as an average-case assumption to deduce computational hardness of detection of satisfiability formulas with lightly planted solutions (Berthet and Ellenberg, 2015). Additionally, several unconditional computational hardness results are shown for statistical problems under constraints of learning models (Feldman et al., 2013; Feldman, Perkins and Vempala, 2015). The present work has two main differences compared to previous computational lower bound results. First, in a detection setting, these lower bounds concern two specific distributions (for the null and alternative hypothesis), while ours is valid for all sub-Gaussian distributions, and there is no alternative distribution. Secondly, our result is not based on the usual assumption for the Planted Clique problem. Instead, we use a weaker assumption on a problem of detecting planted dense graphs. This does not mean that the planted graph is a random graph with edge probability $q > 1/2$ as considered in (Arias-Castro and Verzelen, 2013; Bhaskara et al., 2010; Awasthi et al., 2015), but that it can be *any graph* with an unexpectedly high number of edges (see section 3.4.1). This choice is made to strengthen our result: it would 'survive' the discovery of an algorithm that would use very specific properties of cliques (or even of random dense graphs) to detect their presence. As a consequence, the analysis of our reduction is more technically complicated.

This chapter is organised in the following manner. We recall in Section 3.2 the definition of the restricted isometry property, and some of its known properties. In Section 3.3, we define the notion of certifier, and prove the existence of a computationally efficient certifier in a sub-optimal regime. Our main result is developed in Section 3.4, focused on the hardness of average-case certification. The proofs of the main results are in Appendix 3.5 and those of auxiliary results in Appendix 3.6.

3.2 Restricted isometry property

3.2.1 Formulation

We use the definition of Candès and Tao (2005), who introduced the notion of restricted isometry. Below, for a vector $u \in \mathbb{R}^p$, we define $\|u\|_0$ to be the number of its non-zero entries.

Definition 3.1 (RIP)**.** A matrix $X \in \mathbb{R}^{n\times p}$ satisfies the *restricted isometry property* with sparsity $k \in \{1, \dots, p\}$ and distortion $\theta \in (0, 1)$, denoted by $X \in \mathrm{RIP}_{n,p}(k, \theta)$, if it holds that

$$1 - \theta \leq \|Xu\|_2^2 \leq 1 + \theta,$$

for every $u \in \mathbb{S}^{p-1}(k) := \{u \in \mathbb{R}^p : \|u\|_2 = 1, \|u\|_0 \leq k\}$.

This can be equivalently defined by a property on submatrices of the design matrix: X is in $\mathrm{RIP}_{n,p}(k, \theta)$ if and only if for any set S of k columns of X, the submatrix, X_{*S}, formed by taking any these columns is almost an isometry, i.e. if the spectrum of its Gram matrix is contained in the interval $[1 - \theta, 1 + \theta]$:

$$\|X_{*S}^\top X_{*S} - I_k\|_{\mathrm{op}} \leq \theta\,.$$

Denote by $\|\cdot\|_{\mathrm{op},k}$ the *k-sparse operator norm*, defined for a matrix A as

$$\|A\|_{\mathrm{op},k} := \sup_{x \in \mathbb{S}^{p-1}(k)} \|Ax\|_2.$$

This yields another equivalent formulation of the RIP property: $X \in \mathrm{RIP}_{n,p}(k,\theta)$ if and only if

$$\|X^\top X - I_p\|_{\mathrm{op},k} \leq \theta\,.$$

We assume in the following discussion that the distortion parameter θ is upper-bounded by 1. For $v \in \mathbb{R}^p$ and $T \subseteq \{1,\ldots,p\}$, we write v_T for the $\#T$-dimensional vector obtained by restricting v to coordinates indexed by T. Similarly, for an $n \times p$ matrix A and subsets $S \subseteq \{1,\ldots,n\}$ and $T \subseteq \{1,\ldots,p\}$, we write A_{S*} for the submatrix obtained by restricting A to rows indexed by S, A_{*T} for the submatrix obtained by restricting A to columns indexed by T.

3.2.2 Generation via random design

Matrices that satisfy the restricted isometry property have many interesting applications in high-dimensional statistics and compressed sensing. However, there is no known way to generate them deterministically in general. It is even NP-hard to check whether a given matrix X belongs to $\mathrm{RIP}_{n,p}(k,\theta)$ (see, e.g Bandeira et al., 2012). Several deterministic constructions of RIP matrices exist for sparsity level $k \lesssim \theta\sqrt{n}$. For example, using equi-triangular tight frames and Gershgorin's circle theorem, one can construct RIP matrices with sparsity $k \leq \sqrt{n}$ and distortion θ bounded away from 0 (see, e.g. Bandeira et al., 2012). The limitation $k \leq \theta\sqrt{n}$ is known as the 'square root bottleneck'. To date, the only constructions that break the 'square root bottleneck' are due to Bourgain et al. (2011) and Bandeira, Mixon and Moreira (2014), both of which give RIP guarantee for k of order $n^{1/2+\epsilon}$ for some small $\epsilon > 0$ and fixed θ (the latter construction is conditional on a number-theoretic conjecture being true).

Interestingly, though, it is easy to generate large matrices satisfying the restricted isometry property through random design, and compared to the fixed design matrices mentioned in the previous paragraph, these random design constructions are much less restrictive on the sparsity level, typically allowing k up to the order $n/\log p$ (assuming θ is bounded away from zero). They can be constructed easily from any centred sub-Gaussian distribution. We recall that a distribution Q (and its associated random variable) is said to be sub-Gaussian with parameter σ if $\int_{\mathbb{R}} e^{\lambda x}\, dQ(x) \leq e^{\lambda^2\sigma^2/2}$ for all $\lambda \in \mathbb{R}$.

Definition 3.2. Define $\mathcal{Q} = \mathcal{Q}_\sigma$ to be the set of sub-Gaussian distributions Q over $\mathbb{R}$ with zero mean, unit variance, and sub-Gaussian parameter at most σ.

The most common choice for a $Q \in \mathcal{Q}$ is the standard normal distribution $N(0,1)$. Note that by Taylor expansion, for any $Q \in \mathcal{Q}$, we necessarily have $\sigma^2 \geq \int_{\mathbb{R}} x^2\, dQ(x) = 1$. In the rest of the chapter, we treat σ as fixed. Define the normalised distribution $\tilde{Q}$ to be the distribution of $Z/\sqrt{n}$ for $Z \sim Q$. The following well-known result states that by the concentration of measure, random matrices generated with distribution $\tilde{Q}^{\otimes(n\times p)}$ satisfy restricted isometries (see, e.g. Candès and Tao (2005) and Baraniuk et al. (2008)). For completeness, we include a proof that establishes these particular constants stated here. All proofs are deferred to Appendix 3.5 and Appendix 3.6.

Proposition 3.1. *Suppose X is a random matrix with distribution $\tilde{Q}^{\otimes(n\times p)}$, where $Q \in \mathcal{Q}$. It holds that*

$$\mathbb{P}\big(X \in \mathrm{RIP}_{n,p}(k,\theta)\big) \geq 1 - 2\exp\left\{k\log\left(\frac{9ep}{k}\right) - \frac{n\theta^2}{256\sigma^4}\right\}. \tag{3.1}$$

In order to clarify the notion of asymptotic regimes discussed in this chapter, we introduce the following definition.

Definition 3.3. For $0 \leq \alpha \leq 1$, define the asymptotic regime

$$\mathcal{R}_\alpha := \left\{(p_n, k_n, \theta_n)_n : p, k \to \infty \text{ and } n \gg \frac{k_n^{1+\alpha}\log p_n}{\theta_n^2}\right\}.$$

We note that in this notation, Proposition 3.1 implies that for $(p,k,\theta)_n = (p_n,k_n,\theta_n)_n \in \mathcal{R}_0$ we have $\lim_{n\to\infty}\tilde{Q}^{\otimes(n\times p)}(X \in \mathrm{RIP}_{n,p}(k,\theta)) = 1$, and this convergence is uniform over $Q \in \mathcal{Q}$.

3.3 Certification of restricted isometry

3.3.1 Objectives and definition

In practice, it is useful to know with certainty whether a particular realisation of a random design matrix satisfies the RIP condition. It is known that the problem of deciding if a given matrix is RIP is NP-hard (Bandeira et al., 2012). However, NP-hardness is only a statement about worst-case instances. It would still be of great use to have an algorithm that can correctly decide RIP property for an average instance of a design matrix, with some accuracy. Such an algorithm should identify a high proportion of RIP matrices generated through random design and make no false positive claims. We call such an algorithm an *average-case certifier*, or a *certifier* for short.

Definition 3.4 (Certifier)**.** Given a parameter sequence $(p,k,\theta) = (p_n,k_n,\theta_n)$, we define a *certifier for $\tilde{Q}^{\otimes(n\times p)}$-random matrices* to be a sequence $(\psi_n)_n$ of measurable functions $\psi_n : \mathbb{R}^{n\times p} \to \{0,1\}$, such that

$$\psi_n^{-1}(1) \subseteq \mathrm{RIP}_{n,p}(k,\theta) \qquad \text{and} \qquad \limsup_{n\to\infty}\tilde{Q}^{\otimes(n\times p)}\big(\psi_n^{-1}(0)\big) \leq 1/3. \tag{3.2}$$

Note the definition of a certifier depends on both the asymptotic parameter sequence (p_n,k_n,θ_n) and the sub-Gaussian distribution Q. However, when it is clear from the context, we will suppress the dependence and refer to certifiers for $\mathrm{RIP}_{n,p}(k,\theta)$ properties of $\tilde{Q}^{\otimes(n\times p)}$-random matrices simply as 'certifiers'.

The two defining properties in (3.2) can be understood as follows. The first condition means that if a certifier outputs 1, we know with certainty that the matrix is RIP. The second condition means that the certifier is not overly conservative; it is allowed to output 0 for at most one third (with respect to $\tilde{Q}^{\otimes(n\times p)}$ measure) of the matrices. The choice of $1/3$ in the definition of a certifier is made to simplify proofs. However, all subsequent results will still hold if we replace $1/3$ by any constant in $(0,1)$. In view of Proposition 3.1, the second condition in (3.2) can be equivalently stated as

$$\lim_{n\to\infty}\tilde{Q}^{\otimes(n\times p)}\big\{\psi_n(X) = 1 \;\big|\; X \in \mathrm{RIP}_{n,p}(k,\theta)\big\} \geq 2/3.$$

With such a certifier, given an arbitrary problem fitting the sparse linear model, the matrix X could be tested for the restricted isometry property, with some expectation of a positive result.

This would be particularly interesting given a certifier in the parameter regime $n \ll \theta_n^2 k_n^2$, in which presently known polynomial-time certifiers cannot give positive results.

Even though it is not the main focus of the discussion in this chapter, we also note that a certifier ψ with the above properties for some distribution $Q \in \mathcal{Q}$ would form a certifier/distribution couple (ψ, Q), that yields in the usual manner a Las Vegas algorithm to generate RIP matrices. The (random) algorithm keeps generating random matrices $X \sim \tilde{Q}^{\otimes(n\times p)}$ until $\psi_n(X) = 1$. The number of times that the certifier is invoked has a geometric distribution with success probability $\tilde{Q}^{\otimes(n\times p)}\big(\psi_n^{-1}(1)\big)$. Hence, the Las Vegas algorithm runs in randomised polynomial time if and only if ψ_n runs in randomised polynomial time.

3.3.2 Certifier properties

Although our focus is on algorithmically efficient certifiers, we establish first the properties of a certifier that is computationally intractable. This certifier serves as a benchmark for the performance of other candidates. Indeed, we exhibit in the following proposition a certifier, based on the k-sparse operator norm, that works uniformly well in the same asymptotic parameter regime $\mathcal{R}_0$, where $\tilde{Q}^{\otimes(n\times p)}$-random matrices are RIP with asymptotic probability 1. For clarity, we stress that our criterion when judging a certifier will always be its uniform performance over asymptotic regimes $\mathcal{R}_\alpha$ for some $\alpha \in [0, 1]$.

Proposition 3.2. *Suppose $(p, k, \theta) = (p_n, k_n, \theta_n) \in \mathcal{R}_0$. Furthermore, if $Q \in \mathcal{Q}$ and $X \sim \tilde{Q}^{\otimes(n\times p)}$. Then the sequence of tests $(\psi_{op,k})_n$ based on sparse operator norms, defined by*

$$\psi_{op,k}(X) := \mathbb{1}\Big\{ \|X^\top X - I_p\|_{\mathrm{op},k} \le \theta \Big\}.$$

is a certifier for $\tilde{Q}^{\otimes(n\times p)}$-random matrices.

By a direct reduction from the clique problem, one can show that it is NP-hard to compute the k-sparse operator norm of a matrix. Hence the certifier $\psi_{\mathrm{op},k}$ is computationally intractable. The next proposition concerns the certifier property of a test based on the maximum incoherence between columns of the design matrix. It follows directly from a well-known result on the incoherence parameter of a random matrix (see, e.g. Rauhut and Foucart (2013, Proposition 6.2)) and allows the construction of a polynomial-time certifier that works uniformly well in the asymptotic parameter regime $\mathcal{R}_1$.

Proposition 3.3. *Suppose $(p, k, \theta) = (p_n, k_n, \theta_n)$ satisfies $n \ge 196\sigma^4 k^2 \log(p)/\theta^2$. Let $Q \in \mathcal{Q}$ and $X \sim \tilde{Q}^{\otimes(n\times p)}$, then the tests ψ_∞ defined by $\psi_\infty(X) := \mathbb{1}\Big\{ \|X^\top X - I_p\|_\infty \le 14\sigma^2\sqrt{\frac{\log p}{n}} \Big\}$ is a certifier for $\tilde{Q}^{\otimes(n\times p)}$-random matrices.*

Proposition 3.3 shows that, when the sample size n is above $k^2 \log(p)/\theta^2$ in magnitude (in particular, this is satisfied asymptotically when $(p, k, \theta) = (p_n, k_n, \theta_n) \in \mathcal{R}_1$), there is a polynomial time certifier. In other words, in this high-signal regime, the average-case decision problem for RIP property is much more tractable than indicated by the worst-case result. On the other hand, the certifier in Proposition 3.3 works in a much smaller parameter range when compared to $\psi_{\mathrm{op},k}$ in Proposition 3.2. Combining Proposition 3.2 and 3.3, we have the following schematic diagram (Figure 3.3.2). When the sample size is lower than specified in $\mathcal{R}_0$, the property does not hold, with

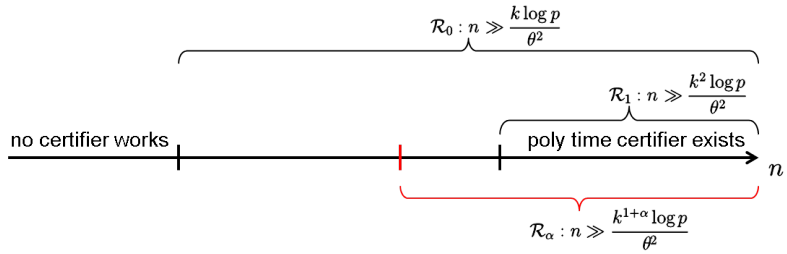

Figure 3.1: Schematic diagram for the existence of certifiers in different asymptotic regimes.

high probability, and no certifier exists. A computationally intractable certifier works uniformly over $\mathcal{R}_0$. On the other end of the spectrum, when the sample size is large enough to be in $\mathcal{R}_1$, a simple certifier based on the maximum incoherence of the design matrix is known to work in polynomial time. This leaves open the question of whether (randomised) polynomial time certifiers can work uniformly well in $\mathcal{R}_0$, or $\mathcal{R}_\alpha$ for any $\alpha \in [0, 1)$. We will see in the next section that, assuming a weaker variant of the Planted Clique hypothesis from computational complexity theory, $\mathcal{R}_1$ is essentially the largest asymptotic regime where a randomised polynomial time certifier can exist.

3.4 Hardness of certification

3.4.1 Planted dense subgraph assumptions

We show in this section that certification of RIP property is an average-case hard problem in the parameter regime $\mathcal{R}_\alpha$ for any $\alpha < 1$. This is precisely the regime not covered by Proposition 3.3. The average-case hardness result is proved via reduction to the planted dense subgraph assumption.

For any integer $m \geq 0$, denote $\mathbb{G}_m$ the collection of all graphs on m vertices. We write $V(G)$ and $E(G)$ for the set of vertices and edges of a graph G. For $H \in \mathbb{G}_\kappa$ where $\kappa \in \{0, \ldots, m\}$, let $\mathcal{G}(m, 1/2, H)$ be the random graph model that generates a random graph G on m vertices as follows. It first picks κ random vertices $K \subseteq V(G)$ and plants an isomorphic copy of H on these κ vertices, then every pair of vertices not in $K \times K$ is connected by an edge independently with probability $1/2$. We write $\mathbf{P}_H$ for the probability measure on $\mathbb{G}_m$ associated with $\mathcal{G}(m, 1/2, H)$. Note that if H is the empty graph, then $\mathcal{G}(m, 1/2, \emptyset)$ describes the Erdős–Rényi random graph. With a slight abuse of notation, we write $\mathbf{P}_0$ in place of $\mathbf{P}_\emptyset$. On the other hand, for $\epsilon \in (0, 1/2]$, if H belongs to the set

$$\mathcal{H} = \mathcal{H}_{\kappa,\epsilon} := \left\{ H \in \mathbb{G}_\kappa : \#E(H) \geq (1/2 + \epsilon) \frac{\kappa(\kappa - 1)}{2} \right\},$$

then $\mathcal{G}(m, 1/2, H)$ generates random graphs that contain elevated local edge density. The planted dense graph problem concerns testing apart the following two hypotheses:

$$H_0 : G \sim \mathcal{G}(m, 1/2, \emptyset) \qquad \text{and} \qquad H_1 : G \sim \mathcal{G}(m, 1/2, H) \text{ for some } H \in \mathcal{H}_{\kappa,\epsilon}. \tag{3.3}$$

It is widely believed that for $\kappa = O(m^{1/2-\delta})$, there does not exist randomised polynomial time tests to distinguish between H_0 and H_1 (see, e.g. Jerrum (1992); Feige and Krauthgamer (2003);

Feldman et al. (2013)). More precisely, we have the following assumption.

Assumption (A1). Fix $\epsilon \in (0, 1/2]$ and $\delta \in (0, 1/2)$. let $(\kappa_m)_m$ be any sequence of integers such that $\kappa_m \to \infty$ and $\kappa_m = O\big(m^{1/2-\delta}\big)$. For any sequence of randomised polynomial time tests $(\phi_m : \mathbb{G}_m \to \{0, 1\})_m$, we have

$$\liminf_m \Big\{ \mathbf{P}_0\big(\phi(G) = 1\big) + \max_{H \in \mathcal{H}_{\kappa,\epsilon}} \mathbf{P}_H\big(\phi(G) = 0)\big) \Big\} > 1/3\,.$$

We remark that if $\epsilon = 1/2$, then $\mathcal{H}_{\kappa,\epsilon}$ contains only the κ-complete graph and the testing problem becomes the well-known planted clique problem (cf. Jerrum (1992) and references in Berthet and Rigollet (2013a,b)).

The difficulty of this problem has been used as a primitive for the hardness of other tasks, such as cryptographic applications, in Juels and Peinado (2000), testing for k-wise dependence in Alon et al. (2007), approximating Nash equilibria in Hazan and Krauthgamer (2011). In this case, Assumption **(A1)** is a version of the planted clique hypothesis (see, e.g. Berthet and Rigollet (2013b, Assumption $\mathbf{A}_{\mathrm{PC}}$) and Assumption **(A1)**(τ) in Chapter 2). We emphasise that Assumption **(A1)** is significantly milder than the planted clique hypothesis (since it allows any $\epsilon \in (0, 1/2]$), or that a hypothesis on planted random graphs. We also note that when $\kappa \geq C_\epsilon \sqrt{m}$, spectral methods can be used to detect such graphs with high probability.

The following theorem relates the hardness of the planted dense subgraph testing problem to the hardness of certifying restricted isometry of random matrices. We recall that the distribution of X is that of an $n \times p$ random matrix with entries independently and identically sampled from $\tilde{Q} \stackrel{d}{=} Q/\sqrt{n}$, for some $Q \in \mathcal{Q}$. We also write Ψ_{rp} for the class of randomised polynomial time certifiers.

Theorem 3.4. *Assume **(A1)** and fix any $\alpha \in [0, 1)$. Then there exists a sequence $(p, k, \theta) = (p_n, k_n, \theta_n) \in \mathcal{R}_\alpha$, such that there is no certifier/distribution couple $(\psi, Q) \in \Psi_{\mathrm{rp}} \times \mathcal{Q}$ with respect to this sequence of parameters.*

Our proof of Theorem 3.4 relies on the following ideas: Given a graph G, an instance of the planted clique problem in the assumed hard regime, we construct n random vectors based on the adjacency matrix of a bipartite subgraph of G, between two random sets of vertices. Each coefficient of these vectors is then randomly drawn from one of two carefully chosen distributions, conditionally on the presence or absence of a particular edge. This construction ensures that if the graph is an Erdős–Rényi random graph (i.e. with no planted graph), the vectors are independent with independent coefficients, with distribution $\tilde{Q}$. Otherwise, we show that with high probability, the presence of an unusually dense subgraph will make it very likely that the matrix does not satisfy the restricted isometry property, for a set of parameters in $\mathcal{R}_\alpha$. As a consequence, if there existed a certifier/distribution couple $(\psi, Q) \in \Psi_{\mathrm{rp}} \times \mathcal{Q}$ in this range of parameters, it could be used - by using as input in the certifier the newly constructed matrix - to determine with high probability the distribution of G, violating our assumption **(A1)**.

We remark that this result holds for *any* distribution in $\mathcal{Q}$, in contrast to computational lower bounds in statistical learning problems, that apply to a specific distribution. For the sake of simplicity, we have kept the coefficients of X identically distributed, but our analysis is not dependent on that fact, and our result can be directly extended to the case where the coefficients are independent, with different distributions in $\mathcal{Q}$.

Theorem 3.4 may be viewed as providing an asymptotic lower bound of the sample size n for the existence of a computationally feasible certifier. It establishes this computational lower bound by exhibiting some specific 'hard' sequences of parameters inside $\mathcal{R}_\alpha$ and shows via a reduction to the planted dense subgraph problem. All hardness results, whether in a worst-case (NP-hardness, or other) or the average-case (by reduction from a hard problem), are by nature statements on the impossibility of accomplishing a task in a computationally efficient manner, uniformly over a range of parameters. They are therefore always based on the construction of a 'hard' sequence of parameters used in the reduction, for which a contradiction is shown. Here, the 'hard' sequence is explicitly constructed in the proof to be some $(p, k, \theta) = (p_n, k_n, \theta_n)$ satisfying $p \geq n$ and $n^{1/(3-\alpha-4\beta)} \ll k \ll n^{1/(2-\beta)-\delta}$, for $\beta \in [0, (1-\alpha)/3)$ and any small $\delta > 0$. The tuning parameter β is to allow additional flexibility in choosing these 'hard' sequences. More precisely, using an averaging trick first seen in Ma and Wu (2015), we are able to show that the existence of such 'hard' sequences is not confined only in the sparsity regime $k \ll n^{1/2}$. We note that in all our 'hard' sequences, θ_n must depend on n. An interesting extension is to see if similar computational lower bounds hold when restricted to a subset of $\mathcal{R}_\alpha$ where θ is constant.

3.5 Appendix: Proofs of the main results

Proof of Theorem 3.4. We prove by contradiction. Assume the contrary, that $(\psi_n)_n$ is a polynomial time computable certifier for $\tilde{Q}^{\otimes(n\times p)}$-random matrices. Let ξ denote the median of $\tilde{Q}$. By the definition of the median, there exists a unique decomposition of the probability measure $\tilde{Q}$ as $\tilde{Q} = \frac{1}{2}\tilde{Q}^+ + \frac{1}{2}\tilde{Q}^-$, where $\tilde{Q}^+$ and $\tilde{Q}^-$ are probability measures supported on $(-\infty, \xi]$ and $[\xi, \infty)$ respectively.

For $\alpha < 1$ and $0 \leq \beta < \frac{1}{3}(1-\alpha)$, let $(p, k, \theta) = (p_n, k_n, \theta_n) \in \mathcal{R}_\alpha$ be a sequence satisfying $p \geq n$, $n^{\frac{1}{3-\alpha-4\beta}} \ll k \ll n^{\frac{1}{2-\beta}-\delta}$ for some $\delta > 0$. Let $L := 10$ and $\ell := \lfloor k^\beta \rfloor$. Define $m := L\ell n$ and $\kappa := Lk$. We check that $\kappa^2 \asymp k^{2-\beta}k^\beta \ll n^{1-\delta}\ell \approx m^{1-\delta'}$ for some positive δ' that depends on δ only. We prove below that Algorithm 3.1, which runs in randomised polynomial time, can distinguish between $\mathbf{P}_0$ and $\mathbf{P}_H$ with zero asymptotic error for any choice of $H \in \mathcal{H}_{\kappa,\epsilon}$.

Algorithm 3.1: Pseudo-code for an algorithm to distinguish between $\mathbf{P}_0$ and $\mathbf{P}_H$.

Input: $m \in \mathbb{N}$, $\kappa \in \{1, \ldots, m\}$, $G \in \mathbb{G}_m$, $L \in \mathbb{N}$

begin

Step 1: Let $N \leftarrow \lfloor m/L \rfloor$, $\ell \leftarrow \lfloor k^\beta \rfloor$, $n \leftarrow \lfloor N/\ell \rfloor$, $p \leftarrow p_n$, $k \leftarrow \lfloor \kappa/L \rfloor$. Draw $u_1, \ldots, u_N, w_1, \ldots, w_N$ uniformly at random without replacement from $V(G)$. Form $A = (A_{ij}) \in \mathbb{R}^{N\times N}$ where $A_{ij} = 2 \cdot \mathbb{1}_{\{u_i \sim w_j\}} - 1$.

Step 2: Let $Y^+ = (Y^+_{ij})$ and $Y^- = (Y^-_{ij})$ be $N \times N$ random matrices independent from all other random variables and from each other, and such that $Y^+_{ij} \overset{\text{i.i.d.}}{\sim} \tilde{Q}^+$ and $Y^-_{ij} \overset{\text{i.i.d.}}{\sim} \tilde{Q}^-$. Define $Z = (Z_{ij})$ by $Z_{ij} \leftarrow \mathbb{1}\{A_{ij} = 1\}Y^+_{ij} + \mathbb{1}\{A_{ij} = -1\}Y^-_{ij}$.

Step 3: For $0 \leq a, b \leq \ell - 1$, define $Z^{(a,b)} \in \mathbb{R}^{n\times n}$ by $Z^{(a,b)}_{i,j} \leftarrow Z_{an+i,bn+j}$. Define $\tilde{X} \leftarrow \ell^{-1}\sum_{0\leq a,b<\ell} Z^{(a,b)}$. Finally, let $X \leftarrow \begin{pmatrix}\tilde{X} & \tilde{X}'\end{pmatrix}$ where $\tilde{X}' \in \mathbb{R}^{n\times(p-n)}$ has entries independently drawn from the distribution $\tilde{Q}$.

Step 4: Let $\phi(G) \leftarrow 1 - \psi_n(X)$.

end

Output: $\phi(G)$

First, we assume that $G \sim \mathbf{P}_0$. Then matrix A from Step 1 of Algorithm 3.1 have independent Rademacher entries, which implies that $X \sim \tilde{Q}^{\otimes(n\times p)}$. Therefore, by (3.2) in Section 3.3 we must have $\limsup \mathbf{P}_0(\phi(G)=1) = \limsup \tilde{Q}^{\otimes(n\times p)}(\psi_n^{-1}(0)) < 1/3$ as desired.

Next, suppose G is generated with probability measure $\mathbf{P}_H$ for some $H \in \mathcal{H}_{\kappa,\epsilon}$. We claim that $\tilde{X} \notin \mathrm{RIP}_{n,n}\Big(k, \frac{ck^2}{n\ell^2}\Big)$ for some absolute positive constant c. By conditions of the theorem, $\frac{k^2}{n\ell^2} \gg \sqrt{\frac{k^{1+\alpha}}{n}} \gg \theta$. Hence if the claim is true, then for large n, $\tilde{X} \notin \mathrm{RIP}_{n,n}(k,\theta)$, which implies that X is not an $\mathrm{RIP}_{n,p}(k,\theta)$ matrix and $\liminf_m \max_{H\in\mathcal{H}_{\kappa,\epsilon}} \mathbf{P}_H\big(\phi(G)=0)\big) < 1/3$, contradicting Assumption **(A1)**.

It remains to verify the claim. Let $K \subseteq V(G)$ be the κ-subset of vertices on which the subgraph H is supported. We write $U = \{u_1,\dots,u_N\}$ and $W = \{w_1,\dots,w_N\}$ for the two random subsets of vertices. Let $N_{U,W;K}$ be the random variable counting the number of edges in G with two endpoints in $U\cap K$ and $W\cap K$ respectively. Then

$$\begin{aligned} N_{U,W;K} &= \#\Big\{\{u,w\} \in E(G) : u \in U\cap K, w \in W \cap K\Big\} \\ &= \sum_{u\in K}\sum_{w\in K} \mathbb{1}\{u\in U\}\mathbb{1}\{w\in W\}\mathbb{1}\{u\sim w\}. \end{aligned}$$

Define

$$\Omega_1 := \left\{N_{U,W;K} \ge \left(\frac{1}{2}+\frac{\epsilon}{4}\right)k^2\right\} \cap \left\{\big|\#U\cap K - k\big| \le \frac{\epsilon}{8}k\right\} \cap \left\{\big|\#W\cap K - k\big| \le \frac{\epsilon}{8}k\right\}.$$

Lemma 3.5 below shows that Ω_1 has asymptotic probability 1. Note Ω_1 is in the σ-algebra of (U,W). Let $U = U_0$ and $W = W_0$ be any realisation satisfying Ω_1. We write $\mathbb{P}^{U_0,W_0}$ and $\mathbb{E}^{U_0,W_0}$ as shorthand for the probability and expectation conditional on $U = U_0$ and $W = W_0$.

For each $j \in \{1,\dots,n\}$, define $s_j := \sum_{u_i\in U\cap K} A_{i,j}$. We write $k_1 := (1-\epsilon/8)k$ and $k_2 := (1+\epsilon/8)k$. Let $S := \{i : u_i \in U\cap K\}$, and let T be a subset of k_1 indices in $\{1,\dots,n\}$ corresponding to the k_1 largest values of s_j (breaking ties arbitrarily). Note that S and T are functions of U and V. On the event $U = U_0$ and $W = W_0$, both $\#S = \#U\cap K$ and $\#W\cap K$ are bounded in the interval $[k_1,k_2]$. In particular, $k_1 \le \#W\cap K$. Therefore, we have

$$\sum_{w_j\in W\cap K} s_j = 2N_{U,W;K} - \#(U\cap K)\times\#(W\cap K) \ge \big\{(1+\epsilon/2) - (1+\epsilon/8)^2\big\}k^2 \ge \frac{\epsilon}{5}k^2.$$

As T indexes columns of A corresponding to largest values of s_js, on the event $\{U = U_0, W = W_0\}$,

$$\sum_{j\in T} s_j \ge \frac{\#T}{\#W\cap K}\frac{\epsilon}{5}k^2 \ge \frac{\epsilon}{5}\frac{k^2k_1}{k_2} \ge \frac{\epsilon}{6}kk_1. \tag{3.4}$$

Define the unit vector $v \in \mathbb{R}^n$ by $v_T := k_1^{-1/2}\mathbf{1}_{k_1}$ and $v_{T^c} := 0$. Note that v is k_1-sparse and hence k-sparse. Conditional on $U = U_0$ and $W = W_0$, we have $Z_{ij} = Y_{ij}^+$ if $A_{ij} = 1$ and $Z_{ij} = Y_{ij}^-$ if $A_{ij} = -1$. By the definitions of $\tilde{Q}^+$ and $\tilde{Q}^-$, and the fact that $\tilde{Q}$ is not a point mass, we have $\mathbb{E}Y_{ij}^+ = -\mathbb{E}Y_{ij}^- = c_1/\sqrt{n}$ for some absolute constant $c_1 > 0$. By (3.4), the sum $\sum_{i\in S, j\in T} Z_{ij}$ can be bounded below in conditional expectation by

$$\mathbb{E}^{U_0,W_0}\sum_{i\in S,j\in T} Z_{ij} \ge \mathbb{E}^{U_0,W_0}\left(\sum_{i\in S,j\in T}(\mathbb{1}\{A_{ij}=1\}Y_{ij}^+ + \mathbb{1}\{A_{ij}=-1\}Y_{ij}^-)\right) \ge \frac{c_1}{\sqrt{n}}\frac{\epsilon}{6}kk_1\,.$$

Lemma 3.7 tells us that both $Y_{ij}^+ - \mathbb{E}Y_{ij}^+$ and $Y_{ij}^- - \mathbb{E}Y_{ij}^-$ are sub-Gaussian with parameters at most $c_2\sigma/\sqrt{n}$ for some absolute constant $c_2 > 0$. By Hoeffding's inequality for sums of sub-Gaussian random variables (see e.g. Vershynin (2012, Proposition 5.10)), we have that

$$\mathbb{P}^{U_0,W_0}\bigg(\sum_{i\in S, j\in T} Z_{ij} > \frac{c_1\epsilon}{12\sqrt{n}}kk_1\bigg) \geq 1 - 2\exp\bigg\{-\frac{(\frac{c_1\epsilon}{12\sqrt{n}}kk_1)^2}{2c_2^2\sigma^2k_1k_2/n}\bigg\} \to 1. \tag{3.5}$$

Combining (3.5) with the fact that $\mathbb{P}(\Omega_1) \to 1$, we deduce that the event $\Omega_2 := \big\{\sum_{i\in S, j\in T} Z_{ij} \geq c_1\epsilon kk_1/(12n^{1/2})\big\}$ has asymptotic probability 1.

Now define $\tilde{S} := \{i \in \{1,\dots,n\} : u_{an+i} \in U \cap K \text{ for some } 0 \leq a \leq \ell-1\}$ and $\tilde{T} := \{j \in \{1,\dots,n\} : w_{bn+j} \in W \cap K \text{ for some } 0 \leq b \leq \ell-1\}$. Also, let $v^{(b)} := (v_{bn+1},\dots,v_{bn+n})^\top$ for $0 \leq b \leq \ell-1$, $\tilde{v}_{\mathrm{sum}} = \sum_{0\leq b\leq \ell-1} v^{(b)}$ and $\tilde{v} := \tilde{v}_{\mathrm{sum}}/\|\tilde{v}_{\mathrm{sum}}\|_2$. By Lemma 3.10, we have that $\|\tilde{v}_{\mathrm{sum}}\|_\infty \leq c_2k_1^{-1/2}$ with asymptotic probability 1, for some c_2 depending on β only. Hence $\|\tilde{v}_{\mathrm{sum}}\|_2 \leq c_2$. Thus, by Cauchy–Schwarz inequality, we have that with asymptotic probability 1,

$$\|\tilde{X}_{\tilde{S}*}\tilde{v}\|_2 \geq \|\tilde{v}_{\mathrm{sum}}\|_2^{-1}(\#\tilde{S})^{-1/2}\|\tilde{X}_{\tilde{S}*}\tilde{v}_{\mathrm{sum}}\|_1 \geq \|\tilde{v}\|_2^{-1}(\#\tilde{S})^{-1/2}\frac{1}{\ell\sqrt{k_1}}\sum_{i\in S, j\in T} Z_{ij} \geq \frac{c_3\epsilon k}{\ell\sqrt{n}}.$$

On the other hand, the submatrix $\tilde{X}_{\tilde{S}^c*}$ has independent and identically distributed entries. By Vershynin (2012, Lemma 5.9), for $i \in \tilde{S}^c$ and $1 \leq j \leq n$, $\tilde{X}_{ij} = \ell^{-1}\sum_{a,b=0}^{\ell-1} Z_{an+i,bn+j}^{(a,b)}$ is a centred sub-Gaussian random variable with sub-Gaussian parameter $\sigma/\sqrt{n}$ and variance $1/n$. Let $\tilde{X}_i$ denote the ith row vector of the matrix $\tilde{X}$, then $\tilde{X}_i^\top\tilde{v}$ is also a centred sub-Gaussian random variable with parameter $\sigma/\sqrt{n}$ and variance $1/n$. Using Lemma 3.9, we derive that

$$\mathbb{P}\bigg(\|\tilde{X}_{S^c*}\tilde{v}\|_2^2 - \frac{n-\#\tilde{S}}{n} \leq -\sqrt{\frac{\log n}{n-\#\tilde{S}}}\bigg) \leq \exp\bigg\{-\frac{\log n}{64\sigma^4}\bigg\} \to 0.$$

Since $\#\tilde{S} \leq k_2$ with asymptotic probability 1, the event $\Omega_3 := \big\{\|\tilde{X}_{\tilde{S}^c*}\tilde{v}\|_2^2 \geq 1 - \frac{k_2}{n} - \sqrt{\frac{2\log n}{n}}\big\}$ has asymptotic probability 1. Finally, since $\tilde{X}\tilde{v} = (\tilde{X}_{\tilde{S}*}\tilde{v}, \tilde{X}_{\tilde{S}^c*}v)^\top$, on $\Omega_2 \cap \Omega_3$,

$$\|\tilde{X}\tilde{v}\|_2^2 = \|\tilde{X}_{\tilde{S}*}\tilde{v}\|_2^2 + \|\tilde{X}_{\tilde{S}^c*}v\|_2^2 \geq 1 + \frac{c_3^2\epsilon^2k^2}{\ell^2 n} - \frac{k_2}{n} - \sqrt{\frac{2\log n}{n}}.$$

The right hand side is at least $1 + ck^2/(n\ell^2)$ for some absolute positive constant c for all large values of n. This verifies the claim and concludes the proof of the theorem. □

Lemma 3.5. *Let G be a graph on m vertices and K a κ-subset of $V(G)$, such that the edge density of G restricted to K is at least $1/2+\epsilon$. Let n, p be integers less than $m/2$. Choose $u_1,\dots,u_n$ and $w_1,\dots,w_p$ independently at random without replacement from $V(G)$. Denote $U = \{u_1,\dots,u_n\}$ and $W = \{w_1,\dots,w_p\}$. Define $N_{U,W;K}$ to be the number of edges with two endpoints in U and W respectively. Then for m, n, p, κ sufficiently large, we have*

$$\mathbb{P}\bigg\{\bigg|\#U\cap K - \frac{n\kappa}{m}\bigg| \geq \frac{\epsilon}{8}\frac{n\kappa}{m}\bigg\} \leq \frac{64m}{\epsilon^2 nk} \qquad \mathbb{P}\bigg\{\bigg|\#W\cap K - \frac{p\kappa}{m}\bigg| \geq \frac{\epsilon}{8}\frac{p\kappa}{m}\bigg\} \leq \frac{64m}{\epsilon^2 pk},$$

and

$$\mathbb{P}\bigg\{N_{U,W;K} \leq \bigg(\frac{1}{2}+\frac{\epsilon}{4}\bigg)\frac{np\kappa^2}{m^2}\bigg\} \leq \frac{16m(p\kappa+n\kappa+m)}{\epsilon^2 np\kappa^2}.$$

Proof. The cardinality of $U \cap K$ has HyperGeom(m, κ, n) distribution. Hence

$$\mathbb{E}(\#U \cap K) = \frac{n\kappa}{m} \quad \text{and} \quad \text{var}(\#U \cap K) = n\frac{\kappa}{m}\frac{m-\kappa}{m}\frac{m-n}{m-1} \leq \frac{n\kappa}{m}.$$

The first inequality in the lemma now follows from an application of Chebyshev's inequality. A similar argument establishes the second inequality. For the final inequality in the lemma, we have that for κ sufficiently large,

$$\begin{aligned}\mathbb{E}(N_{U,W;K}) &= \sum_{u\in K}\sum_{w\in K}\mathbb{P}(u \in U, w \in W)\mathbb{1}\{v \sim w\} \\ &= \frac{np}{m(m-1)}\sum_{u\in K}\sum_{w\in K}\mathbb{1}\{u \sim w\} \geq \Big(\frac{1}{2}+\epsilon\Big)\frac{np\kappa(\kappa-1)}{m(m-1)} \geq \Big(\frac{1}{2}+\frac{\epsilon}{2}\Big)\frac{np\kappa^2}{m^2}..\end{aligned}$$

We then compute the variance of $N_{U,W;K}$ by

$$\begin{aligned}\text{var}(N_{U,W;K}) &= \text{cov}\Big(\sum_{u\in K}\sum_{w\in K}\mathbb{1}\{u \in U, w \in W, u \sim w\}, \sum_{u'\in K}\sum_{w'\in K}\mathbb{1}\{u' \in U, w' \in W, u' \sim w'\}\Big) \\ &= \sum_{u,w,u',w'\in K}\text{cov}\big(\mathbb{1}\{u \in U, w \in W, u \sim w\}, \mathbb{1}\{u' \in U, w' \in W, u' \sim w'\}\big).\end{aligned}$$

We break up the final sum into four terms I, II, III and IV handling sums over subsets of indices $\{(u,w,u',w') \in K^4 : u \neq u', w \neq w'\}$, $\{(u,w,u',w') \in K^4 : u = u', w \neq w'\}$, $\{(u,w,u',w') \in K^4 : u \neq u', w = w'\}$ and $\{(u,w,u',w') \in K^4 : u = u', w = w'\}$ respectively. We bound the four terms separately. For the first term, we have

$$\begin{aligned}\text{I} &= \sum_{u,u',w,w' \text{ distinct}}\Big\{\mathbb{P}(u,u' \in U, w,w' \in W) - \mathbb{P}(u \in U, w \in W)\mathbb{P}(u' \in U, w' \in W)\Big\} \\ &\qquad\qquad \times \mathbb{1}\{v \sim w\}\mathbb{1}\{u' \sim w'\} \\ &= \sum_{u,u',w,w' \text{ distinct}}\left\{\frac{n(n-1)p(p-1)}{m(m-1)(m-2)(m-3)} - \left(\frac{np}{m(m-1)}\right)^2\right\}\mathbb{1}\{u \sim w\}\mathbb{1}\{u' \sim w'\}.\end{aligned}$$

When $m > \max(2n, 2p)$, the term in bracket above is non-positive, hence $\text{I} \leq 0$. For the second term, we get that

$$\begin{aligned}\text{II} &= \sum_{u,w,w' \text{ distinct}}\Big\{\mathbb{P}(u \in U, w,w' \in W) - \mathbb{P}(u \in U, w \in W)\mathbb{P}(u \in U, w' \in W)\Big\} \\ &\qquad\qquad \times \mathbb{1}\{u \sim w\}\mathbb{1}\{u' \sim w'\} \\ &= \sum_{u,w,w' \text{ distinct}}\left\{\frac{np(p-1)}{m(m-1)(m-2)} - \left(\frac{np}{m(m-1)}\right)^2\right\}\mathbb{1}\{u \sim w\}\mathbb{1}\{u \sim w'\} \\ &\leq \frac{np(p-1)}{m(m-1)(m-2)}\sum_{u,w,w' \text{ distinct}}\mathbb{1}\{u \sim w\}\mathbb{1}\{u \sim w'\} \leq \frac{np^2\kappa^3}{m^3}.\end{aligned}$$

Similarly,

$$\text{III} \leq \frac{n(n-1)p\kappa(\kappa-1)(\kappa-2)}{m(m-1)(m-2)} \leq \frac{n^2p\kappa^3}{m^3}.$$

And finally we bound the last term by

$$\text{IV} = \sum_{u,w \text{ distinct}}\Big\{\mathbb{P}(u \in U, w \in W) - \mathbb{P}(u \in U, w \in W)^2\Big\}\mathbb{1}\{u \sim w\} \leq \frac{np\kappa(\kappa-1)}{m(m-1)} \leq \frac{np\kappa^2}{m^2}.$$

Sum up the four terms, we get that

$$\mathrm{var}(N_{U,W;K}) \leq \frac{np\kappa^2}{m^2}\left(\frac{p\kappa}{m} + \frac{n\kappa}{m} + 1\right).$$

An application of Chebyshev's inequality gives that

$$\mathbb{P}\left\{N_{U,W;K} \leq \left(\frac{1}{2} + \frac{\epsilon}{4}\right)\frac{np\kappa^2}{m^2}\right\} \leq \frac{4}{\epsilon}\sqrt{\frac{m(p\kappa + n\kappa + m)}{np\kappa^2}},$$

as desired. □

3.6 Appendix: Ancillary results

Proof of Proposition 3.1. Let X_i denote the ith row vector of X. Then for any fixed $u \in \mathbb{S}^{p-1}(k)$,

$$\mathbb{E}e^{\lambda(X_i^\top u)} = \prod_{1\leq j\leq p} \mathbb{E}e^{\lambda X_{ij}u_j} \leq \prod_j e^{\lambda^2 u_j^2/(2\sigma^2 n)} = e^{\lambda^2/(2\sigma^2 n)}.$$

Applying Lemma 3.9 to $\|Xu\|_2^2 - 1 = n^{-1}\sum_{i=1}^n\{(\sqrt{n}X_i^\top u)^2 - \mathbb{E}(\sqrt{n}X_i^\top u)^2\}$, and using the fact that $\theta/(8\sigma^2) \leq 1$, we have that

$$\mathbb{P}\big(1-\theta \leq \|Xu\|_2^2 \leq 1+\theta\big) \geq 1 - 2e^{-n\theta^2/(64\sigma^4)}.$$

We claim that there is a set $\mathcal{N}$ of cardinality at most $\binom{p}{k}9^k$ such that

$$\sup_{u\in\mathbb{S}^{p-1}(k)} \big|\|Xu\|_2^2 - 1\big| \leq 2\sup_{u\in\mathcal{N}} \big|\|Xu\|_2^2 - 1\big| \tag{3.6}$$

For any cardinality k subset $J \subseteq \{1,\ldots,p\}$, let $B_J = \{u \in \mathbb{S}^{p-1}(k) : u_{J^c} = 0\}$. Each B_J contains a 1/4-net, $\mathcal{N}_J$, of cardinality at most 9^k (Vershynin, 2012, Lemma 5.2). Then $\mathcal{N} := \cup_J \mathcal{N}_J$ form a 1/4-net for $\mathbb{S}^{p-1}(k)$. Define $u_J \in \mathrm{argmax}_{u\in B_J}\|Xu\|^2$ and let v_J be an element in $\mathcal{N}_J$ closest in Euclidean distance to u_J. Define $A := X^\top X - I_p$. We have that

$$|u_J^\top A u_J| \leq |v_J^\top A v_J| + |(u_J - v_J)^\top A v_J| + |u_J^\top A(u_J - v_J)| \leq \max_{u\in\mathcal{N}_I}|u^\top A u|^2 + \frac{1}{2}|u_J^\top A u_J|.$$

Hence

$$\sup_{u\in\mathbb{S}^{p-1}(k)} |u^\top A u| \leq 2\max_{u\in\mathcal{N}}|u^\top A u|,$$

which verifies the claim in (3.6). By a union bound, we obtain that

$$\begin{aligned}\mathbb{P}(X \in \mathrm{RIP}(k,\theta)) &= \mathbb{P}\Big(\sup_{u\in\mathbb{S}^{p-1}(k)}\big|\|Xu\|_2^2 - 1\big| \leq \theta\Big) \geq \mathbb{P}\Big(\sup_{u\in\mathcal{N}}\big|\|Xu\|_2^2 - 1\big| \leq \theta/2\Big)\\ &\geq 1 - 2\binom{p}{k}9^k e^{-n\theta^2/(256\sigma^4)} \geq 1 - 2\exp\left\{k\log\left(\frac{9ep}{k}\right) - \frac{n\theta^2}{256\sigma^4}\right\},\end{aligned}$$

as desired. □

Proof of Proposition 3.2. By definition, $\|X^\top X - I_p\|_{\mathrm{op},k} \leq \theta$ if and only if $X \in \mathrm{RIP}_{n,p}(k,\theta)$. Moreover, by Proposition 3.1, $\tilde{Q}^{\otimes(n\times p)}\{X \in \mathrm{RIP}_{n,p}(k,\theta)\} \to 1$. The proposed test hence satisfies

the two defining properties of a certifier. □

Proof of Proposition 3.3. The proposed test is clearly polynomial time computable (it has time complexity $O(n^2p)$). To verify that it is a certifier, we check that (i) $\psi_n^{-1}(1) \subseteq \mathrm{RIP}_{n,p}(k,\theta)$ and (ii) $\lim_{n\to\infty} \tilde{Q}^{\otimes(n\times p)}(\psi_n^{-1}(1)) > 2/3$.

For (i), on the event $\|X^\top X - I_p\|_\infty \le 14\sigma^2\sqrt{\frac{\log p}{n}}$, for any index set $T \in \{1,\dots,p\}$ of cardinality k, we have that $\|X_{*T}^\top X_{*T} - I_k\|_\infty \le 14\sigma^2\sqrt{\frac{\log p}{n}}$, which implies that

$$\|X_{*T}^\top X_{*T} - I_k\|_{\mathrm{op}} \le 14\sigma^2 k\sqrt{\frac{\log p}{n}} \le \theta$$

as desired. For (ii), we first note that for a general $A \in \mathbb{R}^{p\times p}$

$$\|A\|_\infty = \sup_{S\subseteq\{1,\dots,p\},\#S=2} \|A_{SS}\|_\infty \le \sup_{S\subseteq\{1,\dots,p\},\#S=2} \|A_{SS}\|_{\mathrm{op}} = \|A\|_{\mathrm{op},2}. \tag{3.7}$$

Using Lemma 3.9 and (3.7), we can derive that

$$\begin{aligned}\mathbb{P}\Big\{\|X^\top X - I_p\|_\infty \le 14\sigma^2\sqrt{\frac{\log p}{n}}\Big\} &\ge \mathbb{P}\Big\{\sup_{u\in\mathbb{S}^{p-1}(2)} \big|\|Xu\|_2^2 - 1\big| \le 14\sigma^2\sqrt{\frac{\log p}{n}}\Big\} \\ &\ge 1 - 2\binom{p}{2}9^2\exp\Big\{-\frac{n}{64\sigma^4}\frac{196\sigma^4\log p}{n}\Big\} \ge 1 - 81p^2\exp\{-3\log p\} \to 1.\end{aligned}$$

as desired. □

Lemma 3.6. *Let Z be a non-negative random variable and $r \ge 2$, then $\mathbb{E}(Z^r) \ge \mathbb{E}(|Z - \mathbb{E}Z|^r)$. In other words, centring a nonnegative random variable shrinks its second or higher absolute moments.*

Proof. Let $\mu := \mathbb{E}(Z)$ and define $Y := Z - \mu$. Let P denote the probability measure on R associated with the random variable Y. Hence $\int_{[-\mu,\infty)} y\,dP(y) = 0$. Without loss of generality, we may assume that Z is not a point mass. Then $\int_{[-\mu,0]}(-y)\,dP(y) = \int_{(0,\infty)} y\,dP(y) = A$ for some $A > 0$. For any measurable function $f : \mathbb{R} \to [0,\infty)$, we may write

$$\begin{aligned}A\int_{[-\mu,\infty)} f(y)\,dP(y) &= \int_{[-\mu,0]}(-v)\,dP(v)\int_{(0,\infty)} f(u)\,dP(u) + \int_{(0,\infty)} u\,dP(u)\int_{[-\mu,0]} f(v)dP(v) \\ &= \int_{u\in(0,\infty)}\int_{v\in[-\mu,0]}\Big(\frac{u}{u-v}f(v) - \frac{v}{u-v}f(u)\Big)(u-v)\,dP(v)\,dP(u). \end{aligned}\tag{3.8}$$

Let (U,V) be a bivariate random vector having probability measure

$$\frac{1}{A}(u-v)\mathbb{1}_{(0,\infty)}(u)\mathbb{1}_{[-\mu,0]}(v)\,dP(u)\,dP(v)$$

on $\mathbb{R}^2$ (that this is a probability measure follows from substituting $f(y) \equiv 1$ in (3.8)). Then (3.8) can be rewritten as

$$\mathbb{E}\big\{f(Y)\big\} = \mathbb{E}\Big\{\frac{U}{U-V}f(V) - \frac{V}{U-V}f(U)\Big\}.$$

Now consider choosing f to be $f_1(y) = |y|^r$ and $f_2(y) = (y+\mu)^r$ respectively in the above equation. Note that for $u \in (0,\infty)$ and $v \in [-\mu,0]$ and $r \ge 2$, we always have

$$uf_2(v) - vf_2(u) \ge -vf_2(u) \ge -v(u-v)^r \ge (-v)^r u + (-v)u^r \ge uf_1(v) - vf_1(u).$$

Therefore,

$$\mathbb{E}(|Y|^m) = \mathbb{E}\Big\{\frac{U}{U-V}f_1(V) - \frac{V}{U-V}f_1(U)\Big\} \le \mathbb{E}\Big\{\frac{U}{U-V}f_2(V) - \frac{V}{U-V}f_2(U)\Big\} = \mathbb{E}(|Y+b|^m),$$

as desired. □

Lemma 3.7. *Suppose X is a sub-Gaussian random variable with parameter σ and median ξ. Let $X^+ := X \mid X \ge \xi$ and $X^- := X \mid X < \xi$. Then $X^+ - \mathbb{E}X^+$ and $X^- - \mathbb{E}X^-$ are both sub-Gaussian with parameters are most $c\sigma$ for some absolute constant c.*

Proof. By Vershynin (2012, Lemma 5.5), X is sub-Gaussian with parameter σ, which implies that $(\mathbb{E}|X|^p)^{1/p} \le c_1\sigma\sqrt{p}$ for some absolute constant c_1. Hence by Lemma 3.6, we have

$$\mathbb{E}\big(\big|X^+ - \mathbb{E}X^+\big|^p\big)^{1/p} \le \big(\mathbb{E}\big|X^+\big|^p\big)^{1/p} = 2\big(\mathbb{E}\big|X\,\mathbb{1}\{X \ge \xi\}\big|^p\big)^{1/p} \le 2c_1\sigma\sqrt{p}.$$

Using Vershynin (2012, Lemma 5.5) again, we have that $X^+ - \mathbb{E}X^+$ is sub-Gaussian with parameter at most $c\sigma$ for some absolute constant c. A similar argument holds for $X^- - \mathbb{E}X^-$. □

Lemma 3.8. *Suppose X is a random variable satisfying $\mathbb{E}e^{\lambda X} \le e^{\sigma^2\lambda^2/2}$ for all $\lambda \in \mathbb{R}$. Define $Y := X^2 - \mathbb{E}X^2$. Then $\mathbb{E}e^{\lambda Y} \le e^{16\sigma^4\lambda^2}$ for all $|\lambda| \le \frac{1}{4\sigma^2}$.*

Proof. By Markov's inequality,

$$\mathbb{P}(|X| \ge t) = \mathbb{P}(X \ge t) + \mathbb{P}(-X \ge t) = e^{-t^2/\sigma^2}\mathbb{E}\big(e^{tX/\sigma^2}\big) + e^{-t^2/\sigma^2}\mathbb{E}\big(e^{-tX/\sigma^2}\big) \le 2e^{-t^2/(2\sigma^2)}.$$

From Lemma 3.6, for $r \ge 2$

$$\mathbb{E}(|Y|^r) \le \mathbb{E}(|X|^{2r}) = \int_0^\infty \mathbb{P}(|X| \ge t)(2r)t^{2r-1}\,dt \le \int_0^\infty 4rt^{2r-1}e^{-t^2/(2\sigma^2)}\,dt = 2(2\sigma^2)^r\Gamma(r+1).$$

Consequently, if $|2\sigma^2\lambda| \le 1/2$, then

$$\mathbb{E}e^{\lambda Y} = \sum_{r=0}^\infty \frac{\lambda^r\mathbb{E}Y^r}{r!} \le 1 + 2\sum_{r=2}^\infty (2\sigma^2\lambda)^r \le 1 + 16\sigma^4\lambda^2 \le e^{16\sigma^4\lambda^2},$$

as desired. □

Lemma 3.9. *Let $X_1, X_2, \ldots, X_n$ be independent sub-Gaussian random variables with sub-Gaussian parameters at most σ. Let $Y_i := X_i^2 - \mathbb{E}X_i^2$. Then*

$$\mathbb{P}\Big(\sum_{i=1}^n Y_i \ge \theta\Big) \le \exp\Big\{-\Big(\frac{\theta^2}{64n\sigma^4} \wedge \frac{\theta}{8\sigma^2}\Big)\Big\} \quad \textit{and} \quad \mathbb{P}\Big(\sum_{i=1}^n Y_i \le -\theta\Big) \le \exp\Big\{-\frac{\theta^2}{64n\sigma^4}\Big\}$$

Proof. Using Markov's inequality, we have

$$\mathbb{P}\Big(\sum_{i=1}^n Y_i \ge \theta\Big) = \mathbb{P}\Big(e^{\lambda\sum_i Y_i} \ge e^{\lambda\theta}\Big) \le e^{-\lambda\theta}\prod_i \mathbb{E}e^{\lambda Y_i}.$$

Set $\lambda = \frac{\theta}{32n\sigma^4} \wedge \frac{1}{4\sigma^2}$. By Lemma 3.8, we have

$$\mathbb{P}\bigg(\sum_{i=1}^n Y_i \geq \theta\bigg) \leq e^{-\lambda\theta + 16\lambda^2 n\sigma^4} \leq e^{-\lambda\theta/2},$$

which establishes the first desired inequality. Applying the same argument with $-Y_i$ in place of Y_i we get

$$\mathbb{P}\bigg(\sum_{i=1}^n Y_i \leq -\theta\bigg) \leq \exp\bigg\{-\bigg(\frac{\theta^2}{64n\sigma^4} \wedge \frac{\theta}{8\sigma^2}\bigg)\bigg\}. \tag{3.9}$$

Taylor expand the moment generating function of X_i around 0, we have $\mathbb{E}X_i^2 \leq \sigma^2$. Hence we may assume $\theta \leq n\sigma^2$. Then we have $\frac{\theta^2}{64n\sigma^4} < \frac{\theta}{8\sigma^2}$, which together with (3.9) implies the desired result. □

Lemma 3.10. *Suppose $n\ell$ balls are arranged in an array of n rows and ℓ columns and k balls ($k < n$) are chosen uniformly at random. Let V_i be the number of chosen balls in row i and $V = (V_1, \ldots, V_n)^\top$. Then*

$$\mathbb{P}\bigg(\|V\|_0 \leq k - \frac{k^2}{2n} - \sqrt{k\log k}\bigg) \leq \frac{1}{k^2}.$$

Moreover, if $k \leq n^\gamma$ for some $\gamma < 1$, then

$$\mathbb{P}\big(\|V\|_\infty \geq a\big) \leq n^{1-a(1-\gamma)}\big(1 - n^{-(1-\gamma)}\big).$$

Proof. Let U_i be the number of balls chosen in row i when balls are drawn with replacement from the array and $U = (U_1, \ldots, U_n)^\top$. Then $\|V\|_0$ is stochastically larger than $\|U\|_0$ and $\|V\|_\infty$ is stochastically smaller than $\|U\|_0$. So it suffices to show the desired inequalities with U replacing V. In the following argument, we consider only drawing with replacement.

Let $\mathcal{X} = \{e_1, \ldots, e_n\}$ where e_i denotes the ith standard basis vector in $\mathbb{R}^n$. For $1 \leq r \leq k$, let X_r be uniformly distributed in $\mathcal{X}$. Then $U \overset{d}{=} \sum_{r=1}^k X_r$. We note that changing the value of any one X_r affects the value of $\|U\|_0$ by at most 1. By McDiarmid's inequality (McDiarmid, 1989), we have that for any $t > 0$,

$$\mathbb{P}\big(\|U\|_0 - \mathbb{E}\|U\|_0 \leq -t\big) \leq e^{-\frac{2t^2}{k}}. \tag{3.10}$$

For $1 \leq i \leq n$. Define $J_i = \mathbb{1}\{\text{no ball is chosen in row } i\}$, then

$$\mathbb{E}\|U\|_0 = n - \sum_{i=1}^n \mathbb{E}J_i = n - n(1-1/n)^k \geq k\bigg(1 - \frac{k}{2n}\bigg).$$

Thus, together with (3.10), we have

$$\mathbb{P}\bigg(\|U\|_0 \leq k - \frac{k^2}{2n} - \sqrt{k\log k}\bigg) \leq \mathbb{P}\bigg(\|U\|_0 - \mathbb{E}\|U\|_0 \leq -\sqrt{k\log k}\bigg) \leq e^{-2\log k} = k^{-2},$$

as desired. For the second inequality, we have by a union bound that

$$\mathbb{P}(\|U\|_\infty \geq a) \leq n\sum_{s=a}^k \binom{k}{s} n^{-s} \leq n\sum_{s=a}^\infty (k/n)^s = n\frac{(k/n)^a}{1-k/n} \leq n^{1-a(1-\gamma)}(1 - n^{-(1-\gamma)}),$$

as desired. □

Chapter 4

High-dimensional changepoint estimation via sparse projection

4.1 Introduction

One of the most commonly-encountered issues with Big Data is heterogeneity. When collecting vast quantities of data, it is usually unrealistic to expect that stylised, traditional statistical models of independent and identically distributed observations can adequately capture the complexity of the underlying data generating mechanism. Departures from such models may take many forms, including missing data, correlated errors and data combined from multiple sources, to mention just a few.

When data are collected over time, heterogeneity often manifests itself through non-stationarity, where the data generating mechanism varies with time. Perhaps the simplest form of non-stationarity assumes that population changes occur at a relatively small number of discrete time points. If correctly estimated, these 'changepoints' can be used to partition the original data set into shorter segments, which can then be analysed using methods designed for stationary time series. Moreover, the locations of these changepoints are often themselves of great practical interest.

In this chapter, we study high-dimensional time series that may have changepoints; moreover, we consider in particular settings where at a changepoint, the mean structure changes in a sparse subset of the coordinates. Despite their simplicity, such models are of great interest in a wide variety of applications. For instance, in the case of stock price data, it may well be the case that stocks in related industry sectors experience virtually simultaneous 'shocks' (Chen and Gupta, 1997). In internet security monitoring, a sudden change in traffic at multiple routers may be an indication of a distributed denial of service attack (Peng, Leckie and Ramamohanarao, 2004). In functional Magnetic Resonance Imaging (fMRI) studies, a rapid change in blood oxygen level dependent (BOLD) contrast in a subset of voxels may suggest neurological activity of interest (Aston and Kirch, 2012).

Our main contribution is to propose a new method for estimating the number and locations of the changepoints in such high-dimensional time series, a challenging task in the absence of knowledge of the coordinates that undergo a change. In brief, we first seek a good projection direction, which should ideally be closely aligned with the vector of mean changes. We can then apply an existing univariate changepoint estimation algorithm to the projected series. For this reason, we

call our algorithm `inspect`, short for informative sparse projection for estimation of changepoints. Software implementing the methodology is available in the `R` package `InspectChangepoint` (Wang and Samworth, 2016b).

In more detail, in the single changepoint case, our first observation is that at the population level, the vector of mean changes is the leading left singular vector of the matrix obtained as the cumulative sum (CUSUM) transformation of the mean matrix of the time series. This motivates us to begin by applying the CUSUM transformation to the time series. Unfortunately, computing the k-sparse leading left singular vector of a matrix is a combinatorial optimisation problem, but nevertheless, we are able to formulate an appropriate convex relaxation of the problem, similar to the semidefinite relaxation (2.5) constructed in Chapter 2. We then derive our projection direction from the optimiser of this convex problem. At the second stage of our algorithm, we compute the vector of CUSUM statistics for the projected series, identifying a changepoint if the maximum absolute value of this vector is sufficiently large. For the case of multiple changepoints, we combine our single changepoint algorithm with the method of Wild Binary Segmentation (Fryzlewicz, 2014) to identify changepoints recursively.

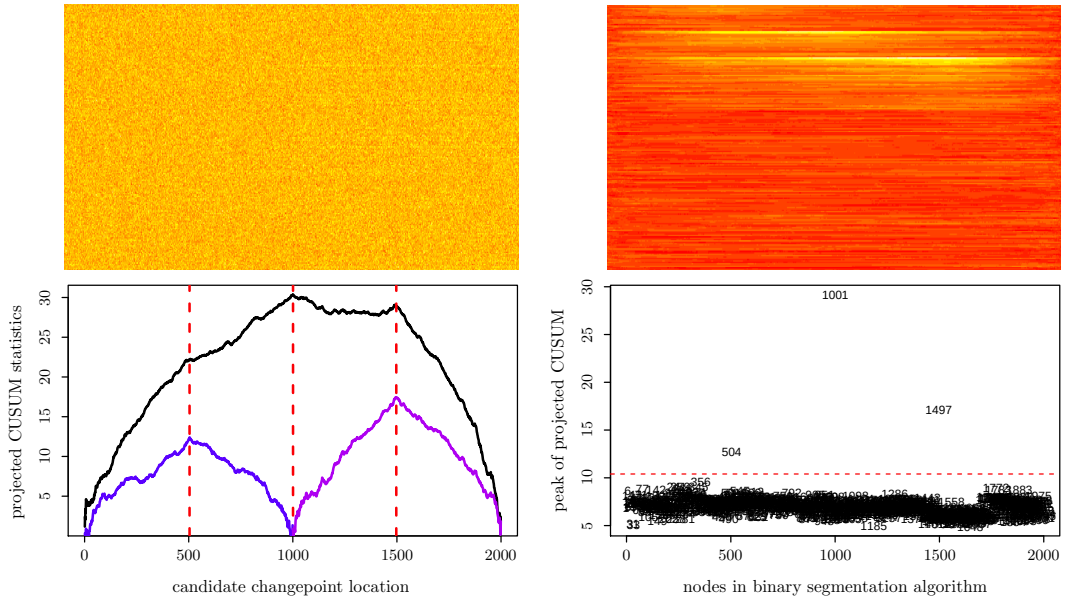

Figure 4.1: An example of `inspect` algorithm in action. Top-left: visualisation of the data matrix. Top-right: its CUSUM transformation. Bottom-left: overlay of the projected CUSUM statistics for the three changepoints detected. Bottom-right: visualisation of thresholding; the three detected changepoints are above the threshold (dotted red line) whereas the remaining numbers are the test statistics obtained if we run the wild binary segmentation to completion without applying a termination criterion.

A brief illustration of the `inspect` algorithm in action is given in Figure 4.1. Here, we simulated a 2000×1000 data matrix having independent normal columns with identity covariance and with three changepoints in the mean structure at locations $500, 1000$ and 1500. Changes occur in 40 coordinates, where consecutive changepoints overlap in half of their coordinates, and the squared ℓ_2 norms of the vectors of mean changes were 0.4, 0.9 and 1.6 respectively. The top-left panel shows the original data matrix and the top-right shows its CUSUM transformation, while the bottom-left panel shows overlays for the three detected changepoints of the univariate CUSUM statistics after

projection. Finally, the bottom-right panel displays the largest absolute values of the projected CUSUM statistics obtained by running the wild binary segmentation algorithm to completion (in practice, we would apply a termination criterion instead, but this is still helpful for illustrative purposes). We see that the three detected changepoints are very close to their true locations, and it is only for these three locations that we obtain a sufficiently large CUSUM statistic to declare a changepoint.

Our theoretical development proceeds first by controlling the angle between the estimated projection direction and the optimal direction, which is given by the normalised vector of mean changes. Under appropriate conditions, this enables us to provide finite-sample bounds which guarantee that with high probability we both recover the correct number of changepoints, and estimate their locations to within a specified accuracy. Our extensive numerical studies indicate that the algorithm performs extremely well in a wide variety of settings.

The study of changepoint problems dates at least back to Page (1955), and has since found applications in many different areas, including genetics (Olshen et al., 2004), disease outbreak watch (Sparks, Keighley and Muscatello, 2010) and aerospace engineering (Henry, Simani and Patton, 2010), in addition to those already mentioned. There is a vast and rapidly growing literature on different methods for changepoint detection and localisation, especially in the univariate problem. Surveys of various methods can be found in Csörgö and Horváth (1997) and Horváth and Rice (2014). In the case of univariate changepoint estimation, state-of-the-art methods include Pruned Exact Linear Time method (PELT) (Killick, Fearnhead and Eckley, 2012), Wild Binary Segmentation (WBS) (Fryzlewicz, 2014) and Simultaneous Multiscale Changepoint Estimator (SMUCE) (Frick, Munk and Sieling, 2014).

Some of the univariate changepoint methodologies have been extended to multivariate settings. Examples include Horváth, Kokoszka and Steinebach (1999), Ombao, Von Sachs and Guo (2005), Aue et al. (2009) and Kirch, Mushal and Ombao (2014). However, there are fewer available tools for high-dimensional changepoint problems, where both the dimension p and the length n of the data stream may be large, and where we may allow a sparsity assumption on the coordinates of change. Bai (2010) investigates the performance of the least squares estimator of a single changepoint in the high-dimensional setting. Zhang et al. (2010), Horváth and Hušková (2012) and Enikeeva and Harchaoui (2014) consider estimators based on ℓ_2 aggregations of CUSUM statistics in all coordinates, but without using any sparsity assumptions. Enikeeva and Harchaoui (2014) also consider a scan statistic that takes sparsity into account. Jirak (2015) considers an ℓ_∞ aggregation of the CUSUM statistics that works well for sparse changepoints. Cho and Fryzlewicz (2015) propose Sparse Binary Segmentation, which also takes sparsity into account and can be viewed as a hard-thresholding of the CUSUM matrix followed by an ℓ_1 aggregation. Cho (2016) proposes a double-CUSUM algorithm that performs a CUSUM transformation along the location axis on the columwise-sorted CUSUM matrix. In a slightly different setting, Lavielle and Teyssiere (2006), Aue et al. (2009), Bücher et al. (2014), Preuß et al. (2015) and Cribben and Yu (2016) deal with changes in cross-covariance. Aston and Kirch (2014) considered the asymptotic efficiency of detecting a single changepoint in a high-dimensional setting, and the oracle projection-based estimator under cross-sectional dependence structure.

The outline of the rest of the paper is as follows. In Section 4.2, we give a formal description of the problem and the class of data generating mechanisms under which our theoretical results hold. Our methodological development in the single changepoint setting is presented in Section 4.3,

and includes theoretical guarantees on both the projection direction and location of the estimated changepoint. Section 4.4 extends these ideas to the case of multiple changepoints with the aid of Wild Binary Segmentation, and our numerical studies are given in Section 4.5.

We conclude this section by introducing some notation used throughout this chapter. For a vector $u = (u_1, \dots, u_M)^\top \in \mathbb{R}^M$, a matrix $A = (A_{ij}) \in \mathbb{R}^{M \times N}$ and for $q \in [1, \infty)$, we write $\|u\|_q := \big(\sum_{i=1}^M |u_i|^q\big)^{1/q}$ and $\|A\|_q := \big(\sum_{i=1}^M \sum_{j=1}^N |A_{ij}|^q\big)^{1/q}$ for their (entrywise) ℓ_q-norms, as well as $\|u\|_\infty := \max_{i=1,\dots,M} |u_i|$ and $\|A\|_\infty := \max_{i=1,\dots,M, j=1,\dots,N} |A_{ij}|$. We write $\|A\|_* := \sum_{i=1}^{\min(M,N)} \sigma_i(A)$ and $\|A\|_{\mathrm{op}} := \max_i \sigma_i(A)$ respectively for the nuclear norm and operator norm of matrix A, where $\sigma_1(A), \dots, \sigma_{\min(M,N)}(A)$ are its singular values. We also write $\|u\|_0 := \sum_{i=1}^M \mathbb{1}_{\{u_i \neq 0\}}$. For $S \subseteq \{1, \dots, M\}$ and $T \subseteq \{1, \dots, N\}$, we write $u_S := (u_i : i \in S)^\top$ and write $M_{S,T}$ for the $|S| \times |T|$ submatrix of A obtained by extracting the rows and columns with indices in S and T respectively. For two matrices $A, B \in \mathbb{R}^{M \times N}$, we denote their trace inner product as $\langle A, B \rangle = \mathrm{tr}(A^\top B)$. For two non-zero vectors $u, v \in \mathbb{R}^p$, we write $\angle(u, v) := \cos^{-1}(\frac{\langle u, v \rangle}{\|u\|_2 \|v\|_2})$ for the acute angle bounded between them. We let $\mathbb{S}^{p-1} := \{x \in \mathbb{R}^p : \|x\|_2 = 1\}$ be the unit Euclidean sphere in $\mathbb{R}^p$, and let $\mathbb{S}^{p-1}(k) := \{x \in \mathbb{S}^{p-1} : \|x\|_0 \leq k\}$.

4.2 Problem description

Let $X_1, \dots, X_n$ be independent p-dimensional random vectors sampled from

$$X_t \sim N_p(\mu_t, \sigma^2 I_p), \qquad 1 \leq t \leq n, \tag{4.1}$$

and combine the observations into a matrix $X = (X_1, \dots, X_n) \in \mathbb{R}^{p \times n}$. We assume that the mean vectors follow a piecewise-constant structure with at most $\nu + 1$ segments. In other words, there exist ν *changepoints*

$$1 \leq z_1 < z_2 < \cdots < z_\nu \leq n - 1$$

such that

$$\mu_{z_i+1} = \cdots = \mu_{z_{i+1}} =: \mu^{(i)}, \qquad \forall\, 0 \leq i \leq \nu,$$

where we adopt the convention that $z_0 := 0$ and $z_{\nu+1} := n$. For $i = 1, \dots, \nu$, write

$$\theta^{(i)} := \mu^{(i)} - \mu^{(i-1)}$$

for the difference in means between consecutive stationary segments. We assume that the changes in mean are sparse in the sense that there exists $k \in \{1, \dots, p\}$ (typically k is much smaller than p) such that $\|\theta^{(i)}\|_0 \leq k$ for each $i = 1, \dots, \nu$.

Our goal is to estimate the set of changepoints $\{z_1, \dots, z_\nu\}$ in the high-dimensional regime, where p may be comparable to, or even larger than, the length n of the series. The signal strength of the estimation problem is determined by the magnitude of mean changes $\{\theta^{(i)} : 1 \leq i \leq \nu\}$ and the run lengths of stationary segments $\{z_{i+1} - z_i : 0 \leq i \leq \nu\}$, whereas the noise is related to the variance σ^2 and the dimensionality p of the observed data points. We let $\mathcal{P}(n, p, k, \nu, \vartheta, \tau, \sigma^2)$ denote the class of distributions of $X = (X_1, \dots, X_n) \in \mathbb{R}^{p \times n}$ with independent columns drawn from (4.1), where the changepoint locations satisfy

$$n^{-1} \min\{z_{i+1} - z_i : 0 \leq i \leq \nu\} \geq \tau,$$

and the magnitudes of mean changes are such that

$$\|\theta^{(i)}\|_2^2 \geq k\vartheta^2, \qquad \forall\, 1 \leq i \leq \nu.$$

Suppose that an estimation procedure outputs $\hat{\nu}$ changepoints located at $1 \leq \hat{z}_1 < \cdots < \hat{z}_{\hat{\nu}} \leq n-1$. Our finite-sample bounds will imply a rate of convergence for `inspect` in an asymptotic setting where $(p, k, \nu, \vartheta, \tau, \sigma^2) = (p_n, k_n, \nu_n, \vartheta_n, \tau_n, \sigma_n^2)$. In this context, we follow the convention in the literature (e.g. Venkatraman, 1992) and say that the procedure is consistent with rate of convergence ρ_n if

$$\sup_{P \in \mathcal{P}(n,p,k,\nu,\vartheta,\tau,\sigma^2)} \mathbb{P}_P\big\{\hat{\nu} = \nu \text{ and } |\hat{z}_i - z_i| \leq n\rho_n \text{ for all } 1 \leq i \leq \nu\big\} \to 1 \tag{4.2}$$

as $n \to \infty$. We remark that consistency as defined above is a rather strong notion, in the sense that it implies convergence in several other natural metrics. For example, if we let

$$d_{\mathrm{H}}(A, B) := \max\Big\{\sup_{a \in A} \inf_{b \in B} |a - b|, \sup_{b \in B} \inf_{a \in A} |a - b|\Big\}$$

denote the Hausdorff distance between non-empty sets A and B on $\mathbb{R}$, then (4.2) implies that with probability tending to 1,

$$\frac{1}{n} d_{\mathrm{H}}\big(\{\hat{z}_i : 1 \leq i \leq \hat{\nu}\}, \{z_i : 1 \leq i \leq \nu\}\big) \leq \rho_n.$$

Similarly, denote the L_1-Wasserstein distance between probability measures P and Q on $\mathbb{R}$ by

$$d_{\mathrm{W}}(P, Q) := \inf_{(U,V) \sim (P,Q)} \mathbb{E}|U - V|$$

where the infimum is taken over all pairs of random variables U and V defined on the same probability space with $U \sim P$ and $V \sim Q$. Then (4.2) also implies that with probability tending to 1,

$$\frac{1}{n} d_{\mathrm{W}}\bigg(\frac{1}{\hat{\nu}} \sum_{i=1}^{\hat{\nu}} \delta_{\hat{z}_i}, \frac{1}{\nu} \sum_{i=1}^{\nu} \delta_{z_i}\bigg) \leq \rho_n,$$

where δ_a denotes a Dirac point mass at a.

4.3 Sparse projection estimator for a single changepoint

We first consider the problem of estimating a single changepoint (i.e. $\nu = 1$) in a high-dimensional time series dataset $X \in \mathbb{R}^{p \times n}$. For simplicity, write $z := z_1$, $\theta = (\theta_1, \ldots, \theta_p)^\top := \theta^{(1)}$ and $\tau := n^{-1} \min\{z, n - z\}$. We seek to aggregate the rows of the data matrix X in an almost optimal way so as to maximise the signal-to-noise ratio, and then locate the changepoint using a one-dimensional procedure. For any $a \in \mathbb{S}^{p-1}$, $a^\top X$ is a one-dimensional time series with

$$a^\top X_t \sim N(a^\top \mu_t, \sigma^2).$$

Hence, the choice $a = \theta / \|\theta\|_2$ maximises the magnitude of the difference in means between the two segments. However, θ is typically unknown in practice, so we should seek a projection direction

that is close to the oracle projection direction $v := \theta/\|\theta\|_2$. Our strategy is to perform sparse singular value decomposition on the CUSUM transformation of X. The method and limit theory of CUSUM statistics in the univariate case can be traced back to Darling and Erdős (1956). For $p \in \mathbb{N}$ and $n \geq 2$, we define the CUSUM transformation $\mathcal{T}_{p,n} : \mathbb{R}^{p\times n} \to \mathbb{R}^{p\times(n-1)}$ by

$$\begin{aligned}[\mathcal{T}_{p,n}(M)]_{j,t} &:= \sqrt{\frac{t(n-t)}{n}}\Big(\frac{1}{n-t}\sum_{r=t+1}^{n} M_{j,r} - \frac{1}{t}\sum_{r=1}^{t} M_{j,r}\Big) \\ &= \sqrt{\frac{n}{t(n-t)}}\Big(\frac{t}{n}\sum_{r=1}^{n} M_{j,r} - \sum_{r=1}^{t} M_{j,r}\Big). \end{aligned} \tag{4.3}$$

In fact, to simplify the notation, we will write $\mathcal{T}$ for $\mathcal{T}_{p,n}$, since p and n can be inferred from the dimensions of the argument of $\mathcal{T}$. Note also that $\mathcal{T}$ reduces to computing the vector of classical one-dimensional CUSUM statistics when $p = 1$. We write

$$X = \boldsymbol{\mu} + W,$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n) \in \mathbb{R}^{p\times n}$ and $W = (W_1, \ldots, W_n)$ is a $p \times n$ random matrix with independent $N_p(0, \sigma^2 I_p)$ columns. Let $T := \mathcal{T}(X)$, $A := \mathcal{T}(\boldsymbol{\mu})$ and $E := \mathcal{T}(W)$, so by the linearity of the CUSUM transformation we have the decomposition

$$T = A + E.$$

In the single changepoint case, the entries of the matrix A can be computed explicitly:

$$A_{j,t} = \begin{cases} \sqrt{\frac{t}{n(n-t)}}(n-z)\theta_j, & \text{if } t \leq z \\ \sqrt{\frac{n-t}{nt}}z\theta_j, & \text{if } t > z. \end{cases}$$

Hence we can write

$$A = \theta\gamma^\top, \tag{4.4}$$

where

$$\gamma := \frac{1}{\sqrt{n}}\Big(\sqrt{\frac{1}{n-1}}(n-z), \sqrt{\frac{2}{n-2}}(n-z), \ldots, \sqrt{z(n-z)}, \sqrt{\frac{n-z-1}{z+1}}z, \ldots, \sqrt{\frac{1}{n-1}}z\Big)^\top. \tag{4.5}$$

In particular, this implies that the oracle projection direction is the leading left singular vector of the rank 1 matrix A. In the ideal case where k is known, we could in principle let $\hat{v}_{\max,k}$ be a k-sparse leading left singular vector of T, defined by

$$\hat{v}_{\max,k} \in \operatorname*{argmax}_{\tilde{v}\in\mathbb{S}^{p-1}(k)} \|T^\top\tilde{v}\|_2, \tag{4.6}$$

and it can then be shown using a perturbation argument akin to the Davis–Kahan 'sin θ' theorem (cf. Davis and Kahan (1970) and Chapter 1) that $\hat{v}_{\max,k}$ is a consistent estimator of the oracle projection direction v under mild conditions (see Proposition 4.9 in Section 4.7). However, the optimisation problem in (4.6) is non-convex and hard to implement. In fact, computing the k-sparse leading left singular vector of a matrix is known to be NP-hard (e.g. Tillmann and Pfetsch (2014)). The naive algorithm that scans through all possible k-subsets of the rows of T has running

time exponential in k, which quickly becomes impractical to run for even moderate sizes of k.

A natural approach to remedy this computational issue is to work with a convex relaxation of the optimisation problem (4.6) instead. In fact, we can write

$$\begin{aligned}\max_{u\in\mathbb{S}^{p-1}(k)} \|u^\top T\|_2 &= \max_{u\in\mathbb{S}^{p-1}(k), w\in\mathbb{S}^{n-2}} u^\top T w \\ &= \max_{u\in\mathbb{S}^{p-1}, w\in\mathbb{S}^{n-2}, \|u\|_0\le k} \langle uw^\top, T\rangle = \max_{M\in\mathcal{M}} \langle M, T\rangle, \end{aligned} \tag{4.7}$$

where $\mathcal{M} := \{M \in \mathbb{R}^{p\times(n-1)} : \|M\|_* = 1, \mathrm{rank}(M) = 1, M \text{ has at most } k \text{ non-zero rows}\}$. The final expression in (4.7) has a convex (linear) objective function $M \mapsto \langle M, T\rangle$. The requirement $\mathrm{rank}(M) = 1$ in the constraint set $\mathcal{M}$ is equivalent to $\|\sigma(M)\|_0 = 1$, where $\sigma(M) := (\sigma_1(M), \ldots, \sigma_{\min(p,n-1)}(M))^\top$ is the vector of singular values of M. This motivates us to absorb the rank constraint into the nuclear norm constraint, which we relax from an equality constraint to an inequality constraint in order to make it convex. Furthermore, we can relax the row sparsity constraint in the definition of $\mathcal{M}$ to an entrywise ℓ_1-norm penalty. The optimisation problem of finding

$$\hat{M} \in \operatorname*{argmax}_{M\in\mathcal{S}_1} \big\{\langle T, M\rangle - \lambda\|M\|_1\big\}, \tag{4.8}$$

where $\mathcal{S}_1 := \{M \in \mathbb{R}^{p\times(n-1)} : \|M\|_* \le 1\}$ and $\lambda > 0$ is a tuning parameter to be chosen later, is therefore a convex relaxation of (4.6). The convex problem (4.8) may be solved using the alternating direction method of multipliers algorithm (ADMM, see e.g. Gabay and Mercier (1976); Boyd et al. (2011)) as in Algorithm 4.1. More specifically, by a variable-splitting trick, the optimisation

Algorithm 4.1: Pseudo-code for an ADMM algorithm that computes the solution to the optimisation problem (4.8).

Input: $T \in \mathbb{R}^{p\times(n-1)}$, $\lambda > 0$.
Set: $Y = Z = R = \mathbf{0} \in \mathbb{R}^{p\times(n-1)}$
repeat
  $Y \leftarrow \Pi_{\mathcal{S}_1}(Z - R + T)$
  $Z \leftarrow \mathbf{soft}(Y + R, \lambda)$
  $R \leftarrow R + (Y - Z)$
until $Y - Z$ *converges to* 0
$\hat{M} \leftarrow Y$
Output: $\hat{M}$

problem in (4.8) is equivalent to maximising $\langle T, Y\rangle - \lambda\|Z\|_1 - \mathbb{I}_{\mathcal{S}_1}(Y)$ subject to $Y = Z$, where $\mathbb{I}_{\mathcal{S}_1}$ is the function that is 0 on $\mathcal{S}_1$ and ∞ on $\mathcal{S}_1^c$. Its augmented Lagrangian is given by

$$L(Y, Z, R) := \langle T, Y\rangle - \mathbb{I}_{\mathcal{S}_1}(Y) - \lambda\|Z\|_1 - \langle R, Y - Z\rangle - \frac{1}{2}\|Y - Z\|_2^2,$$

with the Lagrange multiplier R being the dual variable. Each iteration of the main loop in Algorithm 4.1 first performs a primal update by maximising $L(Y, Z, R)$ marginally with respect to Y and Z, then followed by a dual gradient update of R with constant step size. The function $\Pi_{\mathcal{S}_1}(\cdot)$ in Algorithm 4.1 denotes projection onto the convex set $\mathcal{S}_1$ with respect to the Frobenius norm distance. If $A = UDV^\top$ is the singular value decomposition of $A \in \mathbb{R}^{p\times(n-1)}$ with $\mathrm{rank}(A) = r$, where D is a diagonal matrix with diagonal entries $d_1, \ldots, d_r$, then $\Pi_{\mathcal{S}_1}(A) = U\tilde{D}V^\top$, where $\tilde{D}$ is a diagonal matrix with entries $\tilde{d}_1, \ldots, \tilde{d}_r$ such that $(\tilde{d}_1, \ldots, \tilde{d}_r)^\top$ is the Euclidean projection of the

vector $(d_1, \ldots, d_r)^\top$ onto the standard $(r-1)$-simplex

$$\Delta^{r-1} := \Big\{ (x_1, \ldots, x_r)^\top \in \mathbb{R}^r : \sum_{\ell=1}^{r} x_\ell = 1 \text{ and } x_\ell \geq 0 \text{ for all } \ell \Big\}.$$

For an efficient algorithm for such simplicial projection, see Chen and Ye (2011). The **soft** function in Algorithm 4.1 denotes an entrywise soft-thresholding operator defined by $\big(\mathbf{soft}(A, \lambda)\big)_{ij} = \mathrm{sgn}(A_{ij}) \max\{|A_{ij}| - \lambda, 0\}$ for any $\lambda \geq 0$ and matrix $A = (A_{ij})$.

We remark that one may be interested to further relax (4.8) by replacing $\mathcal{S}_1$ with the larger set $\mathcal{S}_2 := \{M \in \mathbb{R}^{p \times (n-1)} : \|M\|_2 \leq 1\}$. We see from Lemma 4.10 in Section 4.7 that the smoothness of $\mathcal{S}_2$ results in a simple dual formulation, which implies that

$$\tilde{M} := \frac{\mathbf{soft}(T, \lambda)}{\|\mathbf{soft}(T, \lambda)\|_2} = \operatorname*{argmax}_{M \in \mathcal{S}_2} \big\{ \langle T, M \rangle - \lambda \|M\|_1 \big\} \tag{4.9}$$

is the unique optimiser of the primal problem. The soft-thresholding operation is significantly faster than the ADMM algorithm in Algorithm 4.1. Hence by enlarging $\mathcal{S}_1$ to $\mathcal{S}_2$, we can significantly speed up the running time of the algorithm in exchange for some loss in statistical efficiency caused by the further relaxation of the constraint set. See Section 4.5 for further discussion.

Let $\hat{v}$ be the leading left singular vector of

$$\hat{M} \in \operatorname*{argmax}_{M \in \mathcal{S}} \big\{ \langle T, M \rangle - \lambda \|M\|_1 \big\}, \tag{4.10}$$

for either $\mathcal{S} = \mathcal{S}_1$ or $\mathcal{S} = \mathcal{S}_2$. In Proposition 4.1 below, we provide an error bound on $\hat{v}$ as an estimator of the oracle projection direction v. It relies on a generalisation of the curvature lemma in Vu et al. (2013, Lemma 3.1), presented as Lemma 4.4 in Section 4.7.

Proposition 4.1. *Suppose that $\hat{M}$ satisfies (4.10) for either $\mathcal{S} = \mathcal{S}_1$ or $\mathcal{S} = \mathcal{S}_2$. Let $\hat{v} \in \operatorname{argmax}_{\tilde{v} \in \mathbb{S}^{p-1}} \|\hat{M}^\top \tilde{v}\|_2$ be the leading left singular vector of $\hat{M}$. If $n \geq 6$ and if we choose $\lambda \geq 2\sigma\sqrt{\log(p \log n)}$, then*

$$\sup_{P \in \mathcal{P}(n,p,k,1,\vartheta,\tau,\sigma^2)} \mathbb{P}_P \bigg(\sin \angle(\hat{v}, v) > \frac{32\lambda}{\tau\vartheta\sqrt{n}} \bigg) \leq \frac{4}{(p \log n)^{1/2}}.$$

As an illustration, consider $\lambda = 2\sigma\sqrt{\log(p \log n)}$ and the asymptotic regime where $\log p = O(\log n)$, $\vartheta \asymp n^{-a}$ and $\tau \asymp n^{-b}$ for some $a \in \mathbb{R}$ and $b \geq 0$. Then Proposition 4.1 implies that as long as $a + b < 1/2$, we have $\angle(\hat{v}, v) \xrightarrow{\mathrm{p}} 0$.

Algorithm 4.2: Pseudo-code for a single high-dimensional changepoint estimation algorithm.

Input: $X \in \mathbb{R}^{p \times n}$, $\lambda > 0$.
Step 1: Perform the CUSUM transformation $T \leftarrow \mathcal{T}(X)$
Step 2: Use Algorithm 4.1 or (4.9) (with inputs T, λ in either case) to solve for an optimiser $\hat{M}$ of (4.10) for $\mathcal{S} = \mathcal{S}_1$ or $\mathcal{S}_2$
Step 3: Find $\hat{v} \in \operatorname{argmax}_{\tilde{v} \in \mathbb{S}^{p-1}} \|\hat{M}^\top \tilde{v}\|_2$.
Step 4: Let $\hat{z} \in \operatorname{argmax}_{1 \leq t \leq n-1} |\hat{v}^\top T_t|$, where T_t is the tth column of T, and set $\bar{T}_{\max} \leftarrow |\hat{v}^\top T_{\hat{z}}|$
Output: $\hat{z}$, $\bar{T}_{\max}$

After obtaining a good estimator $\hat{v}$ of the oracle projection direction, the natural next step

is to project the data matrix X along the direction $\hat{v}$, and apply an existing one-dimensional changepoint localisation method on the projected data. In this work, we apply a one-dimensional CUSUM transformation to the projected time series and estimate the changepoint by the location of the maximum of the CUSUM vector. Our overall procedure for locating a single changepoint in a high-dimensional time series is given in Algorithm 4.2. In our description of this algorithm, the noise level σ is assumed to be known. If σ is unknown, we can estimate it robustly using, e.g., the median absolute deviation of the marginal one-dimensional time series (Hampel, 1974). Note that for the convenience of later reference, we have required Algorithm 4.2 to output both the estimated changepoint location $\hat{z}$ and the associated maximum absolute post-projection one-dimensional CUSUM statistic $\bar{T}_{\max}$.

From a theoretical point of view, the fact that $\hat{v}$ is estimated using the entire dataset X makes it difficult to analyse the post-projection noise structure. For this reason, in the analysis below, we work with a slight variant of Algorithm 4.2. We assume for convenience that $n = 2n_1$ is even, and define $X^{(1)}, X^{(2)} \in \mathbb{R}^{p\times n_1}$ by

$$X^{(1)}_{j,t} := X_{j,2t-1} \quad \text{and} \quad X^{(2)}_{j,t} := X_{j,2t} \quad \text{for } 1 \le j \le p, 1 \le t \le n_1. \tag{4.11}$$

We then use $X^{(1)}$ to estimate the oracle projection direction and use $X^{(2)}$ to estimate the changepoint location after projection (see Algorithm 4.3). However, in our experience, $\hat{v}$ is almost independent of T and we recommend using Algorithm 4.2 without sample splitting in practice to exploit the full signal strength in the data.

Algorithm 4.3: Pseudo-code for a sample-splitting variant of Algorithm 4.2.

Input: $X \in \mathbb{R}^{p\times n}$, $\lambda > 0$.

Step 1: Perform the CUSUM transformation $T^{(1)} \leftarrow \mathcal{T}(X^{(1)})$ and $T^{(2)} \leftarrow \mathcal{T}(X^{(2)})$.

Step 2: Use Algorithm 4.1 or (4.9) (with inputs $T^{(1)}$, λ in either case) to solve for $\hat{M}^{(1)} \in \operatorname{argmax}_{M\in\mathcal{S}}\{\langle T^{(1)}, M\rangle - \lambda\|M\|_1\}$ with $\mathcal{S} = \{M \in \mathbb{R}^{p\times(n_1-1)} : \|M\|_* \le 1\}$ or $\{M \in \mathbb{R}^{p\times(n_1-1)} : \|M\|_2 \le 1\}$.

Step 3: Find $\hat{v}^{(1)} \in \operatorname{argmax}_{\tilde{v}\in\mathbb{S}^{p-1}} \|(\hat{M}^{(1)})^\top \tilde{v}\|_2$.

Step 4: Let $\hat{z} \in 2\operatorname{argmax}_{1\le t\le n_1-1}\big|(\hat{v}^{(1)})^\top T^{(2)}_t\big|$, where $T^{(2)}_t$ is the tth column of $T^{(2)}$, and set $\bar{T}_{\max} \leftarrow \big|(\hat{v}^{(1)})^\top T^{(2)}_{\hat{z}/2}\big|$.

Output: $\hat{z}, \bar{T}_{\max}$

We summarise the overall estimation performance of Algorithm 4.3 in the following theorem.

Theorem 4.2. *Suppose $\sigma > 0$ is known. Let $\hat{z}$ be the output of Algorithm 4.3 with input X and $\lambda := 2\sigma\sqrt{\log(p\log n)}$. If $n \ge 6$ is even and*

$$\frac{\sigma}{\vartheta\tau}\sqrt{\frac{\log(p\log n)}{n}} \le \frac{\sqrt{3}}{128}, \tag{4.12}$$

then

$$\sup_{P\in\mathcal{P}(n,p,k,1,\vartheta,\tau,\sigma^2)} \mathbb{P}_P\left(\frac{1}{n}|\hat{z} - z| > \frac{32\sigma}{\vartheta\sqrt{k\tau}}\sqrt{\frac{\log n}{n}}\right) \le \frac{4}{\{p\log(n/2)\}^{1/2}} + \frac{2}{n}.$$

Again, to illustrate, suppose we are in the asymptotic regime where $\log p = O(\log n)$, $\vartheta \asymp n^{-a}$, $\tau \asymp n^{-b}$ and $k \asymp n^c$ for some $a \in \mathbb{R}$ and $b \in [0,1]$ and $c \ge 0$. If $a + b < 1/2$, then Theorem 4.2 implies that the output $\hat{z}$ of Algorithm 4.3 is a consistent estimator of the true changepoint z with

rate of convergence $\rho_n = o(n^{-\frac{1-2a-b+c}{2}+\delta})$ for any $\delta > 0$.

4.4 Estimating multiple changepoints

Our algorithm for a single changepoint can be combined with the wild binary segmentation scheme of Fryzlewicz (2014) to sequentially locate multiple changepoints in high-dimensional time series. The principal idea behind a wild binary segmentation procedure is as follows. We first randomly sample a large number of pairs, $(s_1, e_1), \ldots, (s_Q, e_Q)$ uniformly from the set $\{(\ell, r) \in \mathbb{Z}^2 : 0 \leq \ell < r \leq n\}$, and then apply our single changepoint algorithm to $X^{[q]}$, for $1 \leq q \leq Q$, where $X^{[q]}$ is defined to be the submatrix of X obtained by extracting columns $\{s_q + 1, \ldots, e_q\}$ of X. For each $1 \leq q \leq Q$, the single changepoint algorithm (Algorithm 4.2 or 4.3) will estimate an optimal sparse projection direction $\hat{v}^{[q]}$, compute a candidate changepoint location $s_q + \hat{z}^{[q]}$ within the time window $[s_q + 1, e_q]$ and return a maximum absolute CUSUM statistic $\bar{T}^{[q]}_{\max}$ along the projection direction. We aggregate the q candidate changepoint locations by choosing one that maximises the largest projected CUSUM statistic, $T^{[q]}_{\max}$, as our best candidate. If $T^{[q]}_{\max}$ is above a certain threshold value ξ, we admit the best candidate to the set $\hat{Z}$ of estimated changepoint locations and repeat the above procedure recursively on the sub-segments to the left and right of the estimated changepoint. Note that while recursing on a sub-segment, we only consider those time windows that are completely contained in the sub-segment. The precise algorithm is detailed in Algorithm 4.4.

Algorithm 4.4 requires three tuning parameters: a regularisation parameter λ, a Monte Carlo parameter Q for the number of random time windows and a thresholding parameter ξ that determines termination of recursive segmentation. Theorem 4.3 below provides choices for λ, Q and ξ that yield theoretical guarantees for consistent estimation of all changepoints as defined in (4.2).

Algorithm 4.4: Pseudo-code for multiple changepoint algorithm based on sparse singular vector projection and wild binary segmentation.

Input: $X \in \mathbb{R}^{p \times n}$, $\lambda > 0$, $\xi > 0$, $\beta > 0$, $Q \in \mathbb{N}$.

Step 1: Set $\hat{Z} \leftarrow \emptyset$. Draw Q pairs of integers $(s_1, e_1), \ldots, (s_Q, e_Q)$ uniformly at random from the set $\{(\ell, r) \in \mathbb{Z}^2 : 0 \leq \ell < r \leq n\}$.

Step 2: Run `wbs(0, n)` where `wbs` is defined below.

Step 3: Let $\hat{\nu} \leftarrow |\hat{Z}|$ and sort elements of $\hat{Z}$ in increasing order to yield $\hat{z}_1 < \cdots < \hat{z}_{\hat{\nu}}$.

Output: $\hat{z}_1, \ldots, \hat{z}_{\hat{\nu}}$

Function `wbs(`s, e`)`

- Set $\mathcal{Q}_{s,e} \leftarrow \{q : s + n\beta \leq s_q < e_q \leq e - n\beta\}$
- **for** $q \in \mathcal{Q}_{s,e}$ **do**
  - Run Algorithm 4.2 with $X^{[q]}$, λ as input, and let $\hat{z}^{[q]}, \bar{T}^{[q]}_{\max}$ be the output.
- **end**
- Find $q_0 \in \operatorname{argmax}_{q \in \mathcal{Q}_{s,e}} \bar{T}^{[q]}_{\max}$ and set $b \leftarrow s_{q_0} + \hat{z}^{[q_0]}$
- **if** $\bar{T}^{[q_0]}_{\max} > \xi$ **then**
  - $\hat{Z} \leftarrow \hat{Z} \cup \{b\}$
  - `wbs(`s, b`)`
  - `wbs(`b, e`)`
- **end**

end

We remark that if we apply Algorithm 4.2 or 4.3 on the entire dataset X instead of random time

windows of X, and then iterate after segmentation, we arrive at a multiple changepoint algorithm based on the classical binary segmentation scheme. The main disadvantage of this classical binary segmentation procedure is its sensitivity to model misspecification. Algorithms 4.2 and 4.3 are designed to optimise the detection of a single changepoint. When we apply them in conjunction with classical binary segmentation to a time series containing more than one changepoint, the signals from multiple changepoints may cancel each other out in two different ways that will lead to a loss of power. First, as Fryzlewicz (2014) points out in the one-dimensional setting, multiple changepoints may offset each other in CUSUM computation, resulting in a smaller peak of the CUSUM statistic that is more easily contaminated by the noise. Moreover, in a high-dimensional setting, different changepoints can undergo changes in different sets of (sparse) coordinates. This also attenuates the signal strength in the sense that the estimated oracle projection direction from Algorithm 4.1 is aligned to some linear combination of $\theta^{(1)}, \ldots, \theta^{(\nu)}$, but not necessarily well-aligned to any one particular $\theta^{(i)}$. The wild binary segmentation scheme addresses the model misspecification issue by examining sub-intervals of the entire time length. When the number of time windows Q is sufficiently large and τ is not too small, with high probability we have reasonably long time windows that contain each individual changepoint. Hence the single changepoint algorithm will perform well on these segments.

Just as in the case of single changepoint detection, it is easier to analyse the theoretical performance of a sample-splitting version of Algorithm 4.4. However, to avoid notational clutter, we will prove a theoretical result without sample splitting, but with the assumption that whenever Algorithm 4.2 is used within Algorithm 4.4, its second and third steps (i.e. the steps for estimating the oracle projection direction) are carried out on an independent copy X' of X. We refer to such a variant of the algorithm with an access to an independent sample X' as Algorithm 4.4′. Theorem 4.3 below, which proves theoretical guarantees of Algorithm 4.4′, can then be readily adapted to work for a sample-splitting version of Algorithm 4.4, where we replace n by $n/2$ where necessary.

Theorem 4.3. *Suppose $\sigma > 0$ is known and $X, X' \overset{\text{iid}}{\sim} P \in \mathcal{P}(n, p, k, \nu, \vartheta, \tau, \sigma^2)$. Let $\hat{z}_1 < \cdots < \hat{z}_{\hat{\nu}}$ be the output of Algorithm 4.4′ with input X, X', $\lambda := 3\sigma\sqrt{\log(np)}$, $\xi := \lambda$, β and Q. Define $\rho = \rho_n := \frac{200\lambda}{\vartheta\sqrt{kn\tau^3}}$. If $n\tau \geq 14$, $2\rho < \beta < \frac{2}{9}\tau$ and $\rho\sqrt{k\tau} \leq 1$, then*

$$\mathbb{P}_P\big\{\hat{\nu} = \nu \text{ and } |\hat{z}_i - z_i| \leq n\rho \text{ for all } 1 \leq i \leq \nu\big\} \geq 1 - \tau^{-1}e^{-\tau^2 Q/9} - 2n^{-3/2}p^{-5/2}.$$

To illustrate the conditions and conclusion of Theorem 4.3, we again consider the asymptotic setting where $\vartheta \asymp n^{-a}$, $\tau \asymp n^{-b}$, $k \asymp n^c$ and $\log p = O(\log n)$. In this case, the conditions of Theorem 4.3 hold for sufficiently large n if $a + b < 1/2$ and $2a + 5b - c < 1$. When these conditions are satisfied, Theorem 4.3 implies that Algorithm 4.4′ consistently estimates all changepoints with rate of convergence $\rho_n = o(n^{-\frac{1-2a-3b+c}{2}+\delta})$ for any $\delta > 0$.

4.5 Numerical studies

In this section, we examine the empirical performance of the `inspect` algorithm in a range of settings, and compare it with a variety of other recently-proposed methods. In both single- and multiple-changepoint scenarios, the implementation of `inspect` requires the choice of a regularisation parameter $\lambda > 0$ to be used in Algorithm 4.1 (which is called in Algorithms 4.2 and 4.4). In our

experience, the theoretical choices $\lambda = \sigma\sqrt{2\log(p\log n)}$ and $\lambda = 3\sigma\sqrt{\log(np)}$ used in Theorems 4.2 and 4.3 produce consistent estimators as predicted by the theory, but are slightly conservative, and in practice we recommend the choice $\lambda = \sigma\sqrt{2^{-1}\log(p\log n)}$ in both cases. The noise level σ is estimated by concatenating the individual time series into a vector of length np and then computing the median absolute deviation using the scaling constant of 1.48 for the normal distribution (Hampel, 1974).

In Step 2 of Algorithm 4.2, we also have a choice between using $\mathcal{S} = \mathcal{S}_1$ and $\mathcal{S}_2$. The following numerical experiment demonstrates the difference in performance of the algorithm for these two choices. We took $n = 200$, $p = 100$, $k = 10$, with a single changepoint located at $z = 100$. Table 4.1 shows the angles between the oracle projection direction and estimated projection directions using both $\mathcal{S}_1$ and $\mathcal{S}_2$ as the signal level ϑ varies from 0.1 to 1. It can be seen that further relaxation from $\mathcal{S}_1$ to $\mathcal{S}_2$ incurs a relatively low cost in terms of the estimation quality of the projection direction, but it offers great improvement in running time due to the closed-form solution (cf. Lemma 4.10). Thus, even though the use of $\mathcal{S}_1$ remains a viable practical choice for offline data sets of moderate size, we use $\mathcal{S} = \mathcal{S}_2$ in the simulations that follow.

ϑ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$\angle(\hat{v}_{\mathcal{S}_1}, v)$	80.3	63.1	51.6	39.4	28.6	25.8	21.7	19.0	16.7	14.4
$\angle(\hat{v}_{\mathcal{S}_2}, v)$	79.5	63.9	52.9	40.6	30.2	27.3	23.4	20.4	18.0	15.6

Table 4.1: Angles (in degrees) between oracle projection direction v and estimated projection directions $\hat{v}_{\mathcal{S}_1}$ (using $\mathcal{S}_1$) and $\hat{v}_{\mathcal{S}_2}$ (using $\mathcal{S}_2$), for different choices of ϑ. Each reported value is averaged over 100 repetitions. Other simulation parameters: $n = 200$, $p = 100$, $k = 10$, $z = 100$, $\sigma^2 = 1$.

We compare the performance of the `inspect` algorithm with sparsified binary segmentation (`sbs`) (Cho and Fryzlewicz, 2015), the double CUSUM algorithm (`dc`) (Cho, 2016) and a scan statistic-based algorithm (`scan`) derived from the work of Enikeeva and Harchaoui (2014). The latter statistic, rewritten in our notation, is

$$L_{\mathrm{scan}} := \max_{1\le\tilde{z}\le n-1} \max_{1\le\tilde{k}\le p} \frac{\sum_{1\le j\le\tilde{k}}(T^2_{(j),\tilde{z}} - 1)}{\kappa\log\left\{\binom{p}{\tilde{k}}np/\alpha\right\}}, \tag{4.13}$$

where $T^2_{(j),\tilde{z}}$ is the jth largest entry in absolute value in the $\tilde{z}$th column of $T := \mathcal{T}(X)$, and $\alpha = 2$ according to the choice in their paper. We remark that Enikeeva and Harchaoui (2014) primarily concerns the use of L_{scan} to test for the existence of a changepoint. However, the scan statistic can be naturally modified into a changepoint location estimator by modifying the outermost max function in (4.13) to an argmax. It can then be extended a multiple changepoint estimation algorithm via a wild binary segmentation scheme in a similar way to our algorithm. Whenever tuning parameters are required in running these algorithms, we adopt the choices suggested by their authors in the relevant papers. In our simulations, all algorithms were run on the same data matrices, and the estimated changepoints over 100 repetitions were then aggregated and compared.

4.5.1 Single changepoint estimation

All four algorithms in our simulation study are top-down algorithms in the sense that their multiple changepoint procedure is built upon a single changepoint estimation submodule, which is

n	p	k	z	ϑ	`inspect`	`dc`	`sbs`	`scan`
1000	200	10	400	0.18	**32.3**	82.2	99.6	46.2
1000	200	14	400	0.11	**97.2**	274.5	215.7	218.1
1000	200	200	400	0.04	**65.5**	262.3	180.1	156.4
1000	500	10	400	0.18	**48.2**	125.7	181.4	106.1
1000	500	22	400	0.11	**86.9**	240.5	235.5	190.3
1000	500	500	400	0.04	24.5	106.4	96.8	**22.5**
1000	1000	10	400	0.18	**48.6**	118.6	185.4	149.4
1000	1000	32	400	0.11	**58.7**	143.9	171.4	151.3
1000	1000	1000	400	0.04	**10.1**	28.1	42.7	15.1
2000	200	10	800	0.11	**126.3**	327.5	293.9	221.1
2000	200	14	800	0.11	**88.1**	213.7	155.2	121.0
2000	200	200	800	0.04	**57.6**	221.3	155.1	60.9
2000	500	10	800	0.11	**169.9**	348.1	456.0	305.5
2000	500	22	800	0.07	**195.2**	578.4	511.8	535.9
2000	500	500	800	0.04	**21.3**	45.0	62.4	27.0
2000	1000	10	800	0.11	**131.5**	416.4	460.5	397.7
2000	1000	32	800	0.07	**138.4**	441.0	448.6	401.6
2000	1000	1000	800	0.04	**6.7**	30.8	33.7	13.8
1000	200	10	400	0.4	**4.1**	8.4	16.9	5.1
1000	200	14	400	0.25	**7.4**	16.7	31.6	9.4
1000	200	200	400	0.11	**4.4**	15.8	12.7	6.2
1000	500	10	400	0.4	**2.9**	8.9	27.5	4.5
1000	500	22	400	0.25	**4.7**	13.0	20.0	7.2
1000	500	500	400	0.07	**3.7**	13.0	22.0	8.4
1000	1000	10	400	0.4	**3.1**	10.8	30.0	6.1
1000	1000	32	400	0.25	**3.0**	12.0	20.3	6.8
1000	1000	1000	400	0.07	**1.9**	10.1	12.0	4.1
2000	200	10	800	0.25	**7.8**	23.9	45.2	11.6
2000	200	14	800	0.18	**12.1**	44.2	47.7	20.5
2000	200	200	800	0.07	**7.6**	41.6	33.1	17.7
2000	500	10	800	0.25	14.3	28.6	54.4	**14.1**
2000	500	22	800	0.18	**14.5**	33.7	35.4	15.8
2000	500	500	800	0.07	**4.8**	16.2	17.6	9.2
2000	1000	10	800	0.25	**10.5**	29.7	68.3	16.1
2000	1000	32	800	0.18	**6.8**	19.3	39.8	12.6
2000	1000	1000	800	0.07	**1.4**	7.7	13.1	4.5

Table 4.2: Root mean squared error for `inspect`, `dc`, `sbs` and `scan` in single changepoint estimation. The smallest root mean squared error is given in bold.

used to locate recursively all changepoints via a (wild) binary segmentation scheme. It is therefore instructive first to compare their performance in the single changepoint estimation task. Our simulations were run for $n \in \{1000, 2000\}$, $p \in \{200, 500, 1000\}$, $k \in \{10, \lceil p^{1/2} \rceil, p\}$, $z = 0.4n$, $\sigma^2 = 1$ and $\vartheta \in \{1, 0.6, 0.4, 0.25, 0.18, 0.11, 0.07, 0.04\}$, with $\theta = (1, 2^{-1/2}, \ldots, k^{-1/2}, 0, \ldots, 0)^\top \in \mathbb{R}^p$. For definiteness, we let the n columns of X be independent, with the leftmost z columns drawn from $N_p(0, \sigma^2 I_p)$ and the remaining columns drawn from $N_p(\theta, \sigma^2 I_p)$. To avoid the influence of different threshold levels on the performance of the algorithms and to focus solely on their estimation precision, we assume that the existence of a single changepoint is known *a priori* and make all algorithms output their estimate of its location; estimation of the number of changepoints in a multiple-changepoint setting is studied in Section 4.5.2 below. In the interests of brevity, in Table 4.2, we report the root mean squared estimation error for only two values of ϑ, chosen to represent low and high signal-to-noise settings respectively. More precisely, for each choice of (n, p, k), the reported values of ϑ are the largest values in the set $\{1, 0.6, 0.4, 0.25, 0.18, 0.11, 0.07, 0.04\}$ such that the root mean squared error of at least one algorithm is less than $0.1n$ and $0.01n$ respectively. The omitted results are qualitatively similar. We also remark that the three choices for the parameter k correspond to constant/logarithmic sparsity, polynomial sparsity and non-sparse settings respectively. As a graphical illustration, Figure 4.2 displays density estimates of the estimated changepoint location by the different algorithms in two different settings taken from Table 4.2. One difficulty in presenting such estimates with kernel density estimators is the fact that different algorithms would require different choices of bandwidth, and these would need to be locally adaptive, due to the relatively sharp peaks. In order to avoid the choice of bandwidth skewing the visual representation, we therefore use the log-concave maximum likelihood estimators for each method (e.g. Dümbgen and Rufibach, 2009; Cule, Samworth and Stewart, 2010), which is both locally adaptive and tuning-parameter free.

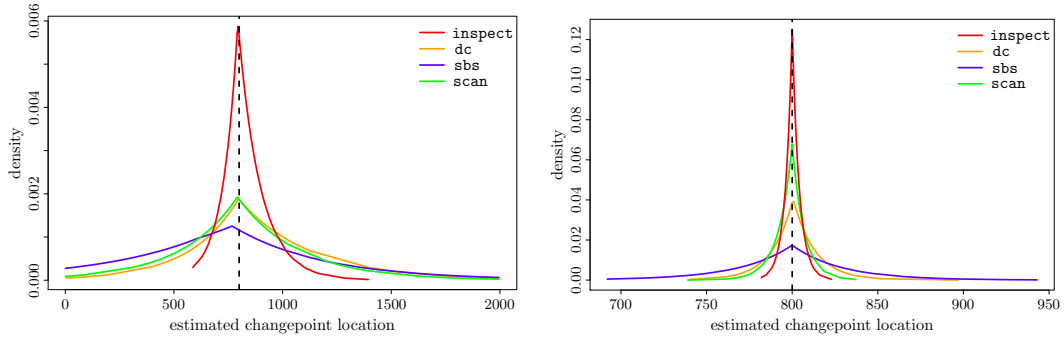

Figure 4.2: Estimated densities of location of changepoint estimates by `inspect`, `dc`, `sbs` and `scan`. Left panel: $(n, p, k, z, \vartheta, \sigma^2) = (2000, 1000, 32, 800, 0.07, 1)$; right panel: $(n, p, k, z, \vartheta, \sigma^2) = (2000, 1000, 32, 800, 0.18, 1)$.

It can be seen from Table 4.2 and Figure 4.2 that `inspect` has extremely competitive performance for the single changepoint estimation task, in both low and high signal-to-noise settings. In particular, despite the fact that it is designed for estimation of sparse changepoints, `inspect` performs relatively well even when $k = p$ (i.e. when the signal is highly non-sparse), especially when the signal strength is relatively large.

We now extend these ideas by investigating empirical performance under several other types of model misspecification. Recall that the noise matrix is $W = (W_{j,t}) := X - \boldsymbol{\mu}$ and we define

$W_1, \ldots, W_n$ to be the column vectors of W. In models $\mathrm{M}_{\mathrm{unif}}$ and $\mathrm{M}_{\mathrm{exp}}$, we replace Gaussian noise by $W_{j,t} \overset{\text{iid}}{\sim} \mathrm{Unif}[-\sqrt{3}\sigma, \sqrt{3}\sigma]$ and $W_{j,t} \overset{\text{iid}}{\sim} \mathrm{Exp}(\sigma) - \sigma$ respectively. In model $\mathrm{M}_{\mathrm{cs,loc}}(\rho)$, we allow the noise to have a short-range cross-sectional dependence by sampling $W_1, \ldots, W_n \overset{\text{iid}}{\sim} N_p(0, \Sigma)$ for $\Sigma := (\rho^{|j-j'|})_{j,j'}$. In model $\mathrm{M}_{\mathrm{cs}}(\rho)$, we extend this to global cross-sectional dependence by sampling $W_1, \ldots, W_n \overset{\text{iid}}{\sim} N_p(0, \Sigma)$ for $\Sigma := (1-\rho)I_p + \frac{\rho}{p}\mathbf{1}_{p\times p}$, where $\mathbf{1}_{p\times p}$ is a $p \times p$ all-one matrix. In model $\mathrm{M}_{\mathrm{temp}}(\rho)$, we consider an auto-regressive AR(1) temporal dependence in the noise by first sampling $W'_{j,t} \overset{\text{iid}}{\sim} N(0, \sigma^2)$ and then setting $W_{j,1} := W'_{j,1}$ and $W_{j,t} := \rho^{1/2} W_{j,t-1} + (1-\rho)^{1/2} W'_{j,t}$ for $2 \leq t \leq n$. We report the performance of the different algorithms in the parameter setting $n = 2000$, $p = 1000$, $k = 32$, $z = 800$, $\vartheta = 0.25$, $\sigma^2 = 1$ in Table 4.3. It can be seen that `inspect` is robust to both temporal and spatial dependence structures, as well as noise misspecification.

Model	n	p	k	z	ϑ	`inspect`	`dc`	`sbs`	`scan`
$\mathrm{M}_{\mathrm{unif}}$	2000	1000	32	800	0.25	**3.0**	13.8	17.6	3.8
$\mathrm{M}_{\mathrm{exp}}$	2000	1000	32	800	0.25	**2.8**	11.9	47.7	5.5
$\mathrm{M}_{\mathrm{cs,loc}}(0.2)$	2000	1000	32	800	0.25	**3.4**	8.4	17.5	6.8
$\mathrm{M}_{\mathrm{cs,loc}}(0.5)$	2000	1000	32	800	0.25	**5.6**	10.8	23.7	8.4
$\mathrm{M}_{\mathrm{cs}}(0.5)$	2000	1000	32	800	0.25	**1.5**	7.5	14.2	3.5
$\mathrm{M}_{\mathrm{cs}}(0.9)$	2000	1000	32	800	0.25	**2.5**	6.5	10.2	2.9
$\mathrm{M}_{\mathrm{temp}}(0.1)$	2000	1000	32	800	0.25	**4.0**	16.9	96.2	10.1
$\mathrm{M}_{\mathrm{temp}}(0.3)$	2000	1000	32	800	0.25	**14.5**	24.9	226.4	14.7

Table 4.3: Root mean squared error for `inspect`, `dc`, `sbs` and `scan` in single changepoint estimation, under different forms of model misspecification.

4.5.2 Multiple changepoint estimation

The use of the 'burn-off' parameter β in Algorithm 4.4 was mainly to facilitate our theoretical analysis. In our simulations, we found that taking $\beta = 0$ rarely resulted in the changepoint being estimated more than once, and we therefore recommend setting $\beta = 0$ in practice, unless prior knowledge of the distribution of the changepoints suggests otherwise. To choose ξ in the multiple changepoint estimation simulation studies, for each (n, p), we first applied `inspect` to 1000 data sets drawn from the null model with no changepoint, and took ξ to be the largest value of $\bar{T}_{\max}$ from Algorithm 4.2. We also set $Q = 1000$.

We consider the simulation setting where $n = 2000$, $p = 200$, $k = 40$, $\sigma^2 = 1$ and $z = (500, 1000, 1500)$. Define $\vartheta^{(i)} := \|\theta^{(i)}\|_2 / \|\theta^{(i)}\|_0^{1/2}$ to be the signal strength at the ith changepoint. We set $(\vartheta^{(1)}, \vartheta^{(2)}, \vartheta^{(3)}) = \vartheta(1, 1.5, 2)$ and let ϑ vary to see the performance of the algorithms at different signal strengths. We also considered different levels of overlap between the coordinates in which the three changes in mean structure occur: in the *complete overlap* case, changes occur in the same k coordinates at each changepoint; in the *half overlap* case, the changes occur in coordinates $\frac{i-1}{2}k + 1, \ldots, \frac{i+1}{2}k$ for $i = 1, 2, 3$; in the *no overlap* case, the changes occur in disjoint sets of coordinates. Table 4.4 summarises the results. We report both the frequency counts of the number of changepoints detected over 100 runs and two quality measures of the location of changepoints. In particular, since changepoint estimation can be viewed as a special case of classification, the quality of the estimated changepoints can be measured by the Adjusted Rand Index (ARI) of the estimated segmentation against the truth (Rand, 1971; Hubert and Arabie, 1985). We report both the ARI and the percentage of runs for which a particular method attains

the largest ARI among the four. Figure 4.3 gives a pictorial representation of the results for one particular collection of parameter settings. Again, we find that the performance of `inspect` is very encouraging on all performance measures.

$(\vartheta^{(1)}, \vartheta^{(2)}, \vartheta^{(3)})$	method	$\hat{\nu}$						ARI	% best
		0	1	2	3	4	5		
$(0.10, 0.15, 0.20)$	`inspect`	0	0	8	65	27	0	0.90	41
	`dc`	0	0	37	61	2	0	0.84	19
	`sbs`	0	0	3	62	30	5	0.88	18
	`scan`	0	0	63	35	2	0	0.80	22
$(0.08, 0.12, 0.16)$	`inspect`	0	0	39	50	11	0	0.78	41
	`dc`	0	1	74	24	1	0	0.73	25
	`sbs`	0	0	34	48	15	3	0.75	18
	`scan`	0	1	95	4	0	0	0.70	18
$(0.06, 0.09, 0.12)$	`inspect`	0	6	61	28	5	0	0.66	40
	`dc`	0	26	72	2	0	0	0.56	18
	`sbs`	0	9	65	27	4	0	0.63	25
	`scan`	0	11	88	1	0	0	0.68	19
$(0.10, 0.15, 0.20)$	`inspect`	0	0	10	73	14	3	0.91	46
	`dc`	0	0	23	63	13	1	0.86	14
	`sbs`	0	0	6	69	22	3	0.85	24
	`scan`	0	0	65	33	2	0	0.79	16
$(0.08, 0.12, 0.16)$	`inspect`	0	0	23	50	22	5	0.82	52
	`dc`	0	0	47	40	12	1	0.76	22
	`sbs`	0	0	30	48	14	8	0.77	20
	`scan`	0	0	94	6	0	0	0.71	7
$(0.06, 0.09, 0.12)$	`inspect`	0	0	48	42	10	0	0.77	55
	`dc`	0	7	66	23	4	0	0.69	18
	`sbs`	0	0	58	36	6	0	0.70	14
	`scan`	0	11	88	1	0	0	0.68	26
$(0.10, 0.15, 0.20)$	`inspect`	0	0	10	74	15	1	0.92	56
	`dc`	0	0	37	57	6	0	0.81	12
	`sbs`	0	0	2	68	28	2	0.86	18
	`scan`	0	0	63	35	2	0	0.78	17
$(0.08, 0.12, 0.16)$	`inspect`	0	0	38	54	8	0	0.81	54
	`dc`	0	0	73	26	1	0	0.69	12
	`sbs`	0	0	26	60	14	0	0.76	26
	`scan`	0	1	89	10	0	0	0.70	9
$(0.06, 0.09, 0.12)$	`inspect`	0	1	66	31	2	0	0.71	52
	`dc`	0	12	78	10	0	0	0.62	17
	`sbs`	0	1	60	30	8	1	0.66	24
	`scan`	0	21	77	2	0	0	0.61	14

Table 4.4: Multiple changepoint simulation results. The top, middle and bottom blocks refer to the complete, half and no overlap settings respectively. Other simulation parameters: $n = 2000$, $p = 200$, $k = 40$, $z = (500, 1000, 1500)$ and $\sigma^2 = 1$.

4.6 Appendix: Proofs of the main results

Proof of Proposition 4.1. We note that the matrix A as defined in Section 4.3 has rank 1, and its only non-zero singular value is $\|\theta\|_2\|\gamma\|_2$. By Proposition 4.6 in Section 4.7, on the event

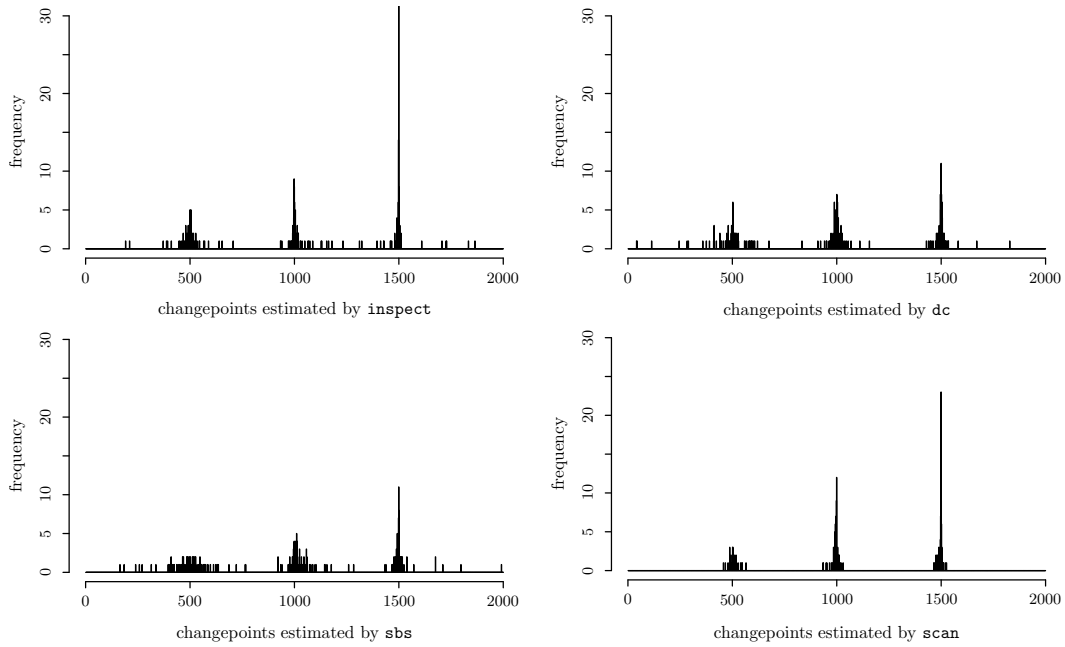

Figure 4.3: Histograms of estimated changepoint locations by `inspect` (top-left), `dc` (top-right), `sbs` (bottom-left) and `scan` (bottom-right) in the half overlap case. Parameter settings: $n = 2000$, $p = 200$, $k = 40$, $z = (500, 1000, 1500)$, $(\vartheta^{(1)}, \vartheta^{(2)}, \vartheta^{(3)}) = (0.10, 0.15, 0.20)$, $\sigma^2 = 1$.

$\Omega_1 := \{\|E\|_\infty \leq \lambda\}$, we have

$$\sin\angle(\hat{v}, v) \leq \frac{8\lambda\sqrt{kn}}{\|\theta\|_2\|\gamma\|_2}.$$

By definition, $\|\theta\|_2 \geq \sqrt{k}\vartheta$, and by Lemma 4.8 in Section 4.7, we have $\|\gamma\|_2 \geq \frac{1}{4}n\tau$. Thus, $\sin\angle(\hat{v}, v) \leq \frac{32\lambda}{\vartheta\tau\sqrt{n}}$ on Ω_1. It remains to verify that $\mathbb{P}(\Omega_1^c) \leq 4(p\log n)^{-1/2}$ for $n \geq 6$. By Lemma 4.5,

$$\begin{aligned}\mathbb{P}\big(\|E\|_\infty \geq 2\sigma\sqrt{\log(p\log n)}\big) &\leq 2\sqrt{\frac{2}{\pi}}p\lceil\log n\rceil\sqrt{\log(p\log n)}\Big\{1 + \frac{1}{\log(p\log n)}\Big\}(p\log n)^{-2} \\ &\leq 6(p\log n)^{-1}\sqrt{\log(p\log n)} \leq 4(p\log n)^{-1/2}, \end{aligned} \tag{4.14}$$

as desired. □

Proof of Theorem 4.2. Recall the definition of $X^{(2)}$ in (4.11) and the definition $T^{(2)} := \mathcal{T}(X^{(2)})$. Define similarly $\boldsymbol{\mu}^{(2)} = (\mu_1^{(2)}, \ldots, \mu_{n_1}^{(2)}) \in \mathbb{R}^{p\times n_1}$ and a random matrix $W^{(2)} = (W_1^{(2)}, \ldots, W_{n_1}^{(2)})$ taking values in $\mathbb{R}^{p\times n_1}$ by $\mu_t^{(2)} := \mu_{2t}$ and $W_t^{(2)} = W_{2t}$; now let $A^{(2)} := \mathcal{T}(\boldsymbol{\mu}^{(2)})$ and $E^{(2)} := \mathcal{T}(W^{(2)})$. Furthermore, we write $\bar{X} := (\hat{v}^{(1)})^\top X^{(2)}$, $\bar{\mu} := (\hat{v}^{(1)})^\top \boldsymbol{\mu}^{(2)}$, $\bar{W} := (\hat{v}^{(1)})^\top W^{(2)}$, $\bar{T} := (\hat{v}^{(1)})^\top T^{(2)}$, $\bar{A} := (\hat{v}^{(1)})^\top A^{(2)}$ and $\bar{E} := (\hat{v}^{(1)})^\top E^{(2)}$ for the one-dimensional projected images (as row vectors) of the corresponding p-dimensional quantities. We note that $\bar{T} = \mathcal{T}(\bar{X})$, $\bar{A} = \mathcal{T}(\bar{\mu})$ and $\bar{E} = \mathcal{T}(\bar{W})$.

Now, conditional on $\hat{v}^{(1)}$, the random variables $\bar{X}_1, \ldots, \bar{X}_{n_1}$ are independent, with

$$\bar{X}_t \mid \hat{v}^{(1)} \sim N(\bar{\mu}_t, \sigma^2),$$

and the row vector $\bar{\mu}$ undergoes a single change at $z^{(2)} := z/2$ with magnitude of change

$$\bar{\theta} := \bar{\mu}_{z^{(2)}+1} - \bar{\mu}_{z^{(2)}} = (\hat{v}^{(1)})^\top \theta.$$

Finally, let $\hat{z}^{(2)} \in \operatorname{argmax}_{1 \le t \le n_1 - 1} |\bar{T}_t|$, so we may assume the first component of the output of the algorithm is $\hat{z} = 2\hat{z}^{(2)}$. Consider the set $\Upsilon := \{\tilde{v} \in \mathbb{S}^{p-1} : \angle(\tilde{v}, v) \le \pi/3\}$. By condition (4.12) in the statement of the theorem and Proposition 4.1,

$$\mathbb{P}(\hat{v}^{(1)} \in \Upsilon) \ge 1 - 4(p \log n_1)^{-1/2}. \tag{4.15}$$

Note that $\hat{v}^{(1)}$ and $W^{(2)}$ are independent, so $\bar{W}$ has independent $N(0, \sigma^2)$ entries. Hence, by Lemma 4.5 in Section 4.7,

$$\mathbb{P}\big(\|\bar{E}\|_\infty \ge 2\sigma\sqrt{\log n_1}\big) \le \sqrt{\frac{2}{\pi}}\lceil \log n_1 \rceil \bigg(2\sqrt{\log n_1} + \frac{1}{\sqrt{\log n_1}}\bigg) n_1^{-2} \le n_1^{-1}. \tag{4.16}$$

Since $\bar{T} = \bar{A} + \bar{E}$, and since $(\bar{A}_t)_t$ and $(\bar{T}_t)_t$ are respectively maximised at $t = z^{(2)}$ and $t = \hat{z}^{(2)}$, we have on the event $\Omega_0 := \big\{\hat{v}^{(1)} \in \Upsilon, \|\bar{E}\|_\infty \le 2\sigma\sqrt{\log n_1}\big\}$ that

$$\begin{aligned}\bar{A}_{z^{(2)}} - \bar{A}_{\hat{z}^{(2)}} &= (\bar{A}_{z^{(2)}} - \bar{T}_{z^{(2)}}) + (\bar{T}_{z^{(2)}} - \bar{T}_{\hat{z}^{(2)}}) + (\bar{T}_{\hat{z}^{(2)}} - \bar{A}_{\hat{z}^{(2)}}) \\ &\le |\bar{A}_{z^{(2)}} - \bar{T}_{z^{(2)}}| + |\bar{T}_{\hat{z}^{(2)}} - \bar{A}_{\hat{z}^{(2)}}| \le 4\sigma\sqrt{\log n_1}.\end{aligned}$$

The row vector $\bar{A}$ has the following explicit form

$$\bar{A}_t = \begin{cases} \sqrt{\frac{t}{n_1(n_1-t)}}(n_1 - z^{(2)})\bar{\theta}, & \text{if } t \le z^{(2)} \\ \sqrt{\frac{n_1 - t}{n_1 t}} z^{(2)} \bar{\theta}, & \text{if } t > z^{(2)}. \end{cases}$$

Hence, by Lemma 4.11, on the event Ω_0,

$$\hat{z}^{(2)} \in \big[z^{(2)} - 2\Delta z^{(2)}, z^{(2)} + 2\Delta(n_1 - z^{(2)})\big],$$

where $\Delta := 4\sigma\bar{\theta}^{-1}\sqrt{\frac{n_1 \log n_1}{z^{(2)}(n_1 - z^{(2)})}}$. Since $\hat{z} = 2\hat{z}^{(2)}$ and $z = 2z^{(2)}$, we have that on Ω_0,

$$\frac{1}{n}|\hat{z} - z| \le 2\Delta \le \frac{8\sqrt{2}\sigma}{\bar{\theta}}\sqrt{\frac{n \log n}{z(n-z)}} \le \frac{16\sigma}{\bar{\theta}\sqrt{\tau}}\sqrt{\frac{\log n}{n}}. \tag{4.17}$$

On the event Ω_0, we have $\bar{\theta} \ge \sqrt{k}\vartheta/2$. We deduce from (4.15), (4.16) and (4.17) that

$$\mathbb{P}\bigg\{\frac{1}{n}|\hat{z} - z| > \frac{32\sigma}{\vartheta\sqrt{k\tau}}\sqrt{\frac{\log n}{n}}\bigg\} \le 4\bigg\{p \log\bigg(\frac{n}{2}\bigg)\bigg\}^{-1/2} + \frac{2}{n},$$

as desired. □

Proof of Theorem 4.3. For $i \in \{0, 1, \ldots, \nu\}$, we define $J_i := \big[z_i + \lceil \frac{z_{i+1} - z_i}{3} \rceil, z_{i+1} - \lceil \frac{z_{i+1} - z_i}{3} \rceil\big]$ and

$$\Omega_1 := \bigcap_{i=1}^{\nu} \bigcup_{q=1}^{Q} \{s_q \in J_{i-1}, e_q \in J_i\}.$$

By a union bound, we have

$$\mathbb{P}(\Omega_1^c) \le \nu\left(1 - \frac{(z_i - z_{i-1} - 2\lceil\frac{z_i - z_{i-1}}{3}\rceil)(z_{i+1} - z_i - 2\lceil\frac{z_{i+1} - z_i}{3}\rceil)}{n(n+1)/2}\right)^Q$$
$$\le \nu\left(1 - \frac{(z_i - z_{i-1})(z_{i+1} - z_i)}{9n^2}\right)^Q \le \tau^{-1}(1 - \tau^2/9)^Q \le \tau^{-1}e^{-\tau^2 Q/9},$$

where the second inequality uses the fact that $n\tau \ge 14$. For any matrix $M \in \mathbb{R}^{p\times n}$ and $1 \le \ell \le r \le n$, we write $M^{[\ell,r]}$ for the submatrix obtained by extracting columns $\{\ell, \ell+1, \ldots, r\}$ of M. Also define $\boldsymbol{\mu}' := \mathbb{E}X' = \boldsymbol{\mu}$ and $W' := X' - \boldsymbol{\mu}'$. Let $\hat{v}^{[\ell,r]}$ be a leading left singular vector of a maximiser of

$$M \mapsto \langle \mathcal{T}(X'^{[\ell,r]}), M\rangle - \lambda\|M\|_1,$$

for $M \in \mathcal{S}$, where $\mathcal{S} = \mathcal{S}_1$ or $\mathcal{S}_2$. For definiteness, we assume both the maximiser and its leading left singular vector are chosen to be the lexicographically smallest possibilities. For $q = 1, \ldots, Q$, we also write $M^{[q]}$ for $M^{[s_q+1,e_q]}$ and $\hat{v}^{[q]}$ for $\hat{v}^{[s_q+1,e_q]}$. Define events

$$\Omega_2 := \bigcap_{1\le\ell<r\le n} \{\|\mathcal{T}(W'^{[\ell,r]})\|_\infty \le \lambda\},$$
$$\Omega_3 := \bigcap_{1\le\ell<r\le n} \{\|(\hat{v}^{[\ell,r]})^\top \mathcal{T}(W^{[\ell,r]})\|_\infty \le \lambda\}.$$

By Lemma 4.5,

$$\mathbb{P}(\Omega_2^c) \le \binom{n}{2}\sqrt{\frac{2}{\pi}}p\lceil\log n\rceil\left(3\sqrt{\log(np)} + \frac{2}{3\sqrt{\log(np)}}\right)(np)^{-9/2} \le n^{-3/2}p^{-5/2}.$$

Also, since $\hat{v}^{[\ell,r]}$ and X are independent, $(\hat{v}^{[\ell,r]})^\top\mathcal{T}(W)$ has the same distribution as $\mathcal{T}(G)$, where G is a row vector of length $r - \ell + 1$ with independent $N(0, \sigma^2)$ entries. So by Lemma 4.5 again, for sufficiently large n,

$$\mathbb{P}(\Omega_3^c) \le \binom{n}{2}\mathbb{P}\{\|\mathcal{T}(G)\|_\infty > \lambda\} \le n^{-3/2}p^{-5/2}.$$

We claim that the desired event $\Omega^* := \{\hat{\nu} = \nu \text{ and } |\hat{z}_i - z_i| \le n\rho \text{ for all } 1 \le i \le \nu\}$ occurs if the following two statements hold every time the function `wbs` is called in Algorithm 4.4′:

(i) There exist unique $i_1, i_2 \in \{0, 1, \ldots, \nu+1\}$ such that $|s - z_{i_1}| \le n\rho$ and $|e - z_{i_2}| \le n\rho$, where (s, e) is the pair of arguments of the `wbs` function call.

(ii) $\bar{T}_{\max}^{[q_0]} > \xi$ if and only if $i_2 - i_1 \ge 2$, where i_1 and i_2 are the indices defined in (i).

To see this, observe that the set of all arguments used in the calls of the function `wbs` is $\hat{Z} \cup \{0, n\}$, so (i) ensures that

$$\max_{\hat{z}\in\hat{Z}\cup\{0,n\}} \min_{i\in\{0,1,\ldots,\nu+1\}} |\hat{z} - z_i| \le n\rho.$$

If $|\hat{z} - z_i| \le n\rho$, we say $\hat{z}$ is 'identified' to z_i. Moreover, each candidate changepoint b identified by the function call `wbs`(s, e) in Algorithm 4.4′ satisfies $\min\{b - s, e - b\} \ge n\beta > 2n\rho$. It follows that different elements of $\hat{Z} \cup \{0, n\}$ cannot be identified to the same z_i, so no element of $\hat{Z}$ is identified to z_0 or $z_{\nu+1}$, and the second part of the event Ω^* holds. It remains to show that each element of $\{z_1, \ldots, z_\nu\}$ is identified by some element of $\hat{Z}$. To see this, note that if z_i is not identified, we can

let (s^*, e^*) be the shortest interval such that $s^* + 1 \le z_i \le e^*$ and such that (s^*, e^*) are a pair of arguments called by the `wbs` function in Algorithm 4.4′. By (i), the two endpoints s^* and e^* are identified to z_{i_1} and z_{i_2} respectively, say, for some $i_1 \le i-1$ and $i_2 \ge i+1$. But then by (ii) a new point b will be added to $\hat{Z}$ and the recursion continues on the pairs (s^*, b) and (b, e^*), contradicting the minimality of the pair (s^*, e^*).

We now prove by induction on the depth of the recursion that on $\Omega_1 \cap \Omega_2 \cap \Omega_3$, statements (i) and (ii) hold every time `wbs` is called in Algorithm 4.4′. The first time `wbs` is called, $s = 0$ and $e = n$, so (i) is satisfied with the unique choice $i_1 = 0$ and $i_2 = \nu + 1$. This proves the base case. Now suppose `wbs` is called with the pair (s, e) satisfying (i), yielding indices $i_1, i_2 \in \{0, 1, \ldots, \nu+1\}$ with $|s - z_{i_1}| \le n\rho$, $|e - z_{i_2}| \le n\rho$. To complete the inductive step, we need to show that (ii) also holds, and if a new changepoint b is detected, then (i) holds for the pairs of arguments (s, b) and (b, e). We have two cases.

Case 1: $i_2 - i_1 = 1$. In this case, $(s + n\beta, e - n\beta]$ contains no changepoint. Since $\xi = \lambda$, on Ω_3 we always have

$$\bar{T}^{[q_0]}_{\max} = \max_{q \in \mathcal{Q}_{s,e}} \|(\hat{v}^{[q]})^\top \mathcal{T}(X^{[q]})\|_\infty \le \xi,$$

so (ii) is satisfied with no additional changepoint detected.

Case 2: $i_2 - i_1 \ge 2$. On the event Ω_1, for any $i^* \in \{i_1+1, \ldots, i_2-1\}$, there exists $q^* \in \{1, \ldots, Q\}$ such that $s_{q^*} \in J_{i^*-1}$ and $e_{q^*} \in J_{i^*}$. Moreover, since $\min\{s_{q^*} - s, e - e_{q^*}\} \ge \lceil n\tau/3 \rceil - n\rho > n\beta$ by the condition on β in the theorem, we have $q^* \in \mathcal{Q}_{s,e}$. Since there is precisely one changepoint within the segment $(s_{q^*}, e_{q^*}]$, the matrix $\mathcal{T}(\boldsymbol{\mu}'^{[q^*]})$ has rank 1; cf. (4.4). On Ω_2, we have $\|\mathcal{T}(W'^{[q^*]})\|_\infty \le \lambda$. Thus, by Proposition 4.6 and Lemma 4.8 in Section 4.7,

$$\sin \angle\big(\hat{v}^{[q^*]}, \theta^{(i^*)}/\|\theta^{(i^*)}\|_2\big) \le \frac{8\lambda\sqrt{k(e_{q^*} - s_{q^*})}}{\|\theta^{(i^*)}\|_2 n\tau/12} \le \frac{96\lambda}{\vartheta\tau\sqrt{n}} = \frac{96}{200}\rho\sqrt{k\tau} \le \frac{12}{25}$$

under the conditions of the theorem. Therefore, recalling the definition of q_0 in Algorithm 4.4′, and on the event Ω_3,

$$\begin{aligned}
\bar{T}^{[q_0]}_{\max} \ge \bar{T}^{[q^*]}_{\max} &= \|(\hat{v}^{[q^*]})^\top \mathcal{T}(X^{[q^*]})\|_\infty \ge \|(\hat{v}^{[q^*]})^\top \mathcal{T}(\boldsymbol{\mu}^{[q^*]})\|_\infty - \|(\hat{v}^{[q^*]})^\top \mathcal{T}(W^{[q^*]})\|_\infty \\
&\ge \big|(\hat{v}^{[q^*]})^\top \theta^{(i^*)}\big| \sqrt{\frac{(z_{i^*} - s_{q^*})(e_{q^*} - z_{i^*})}{e_{q^*} - s_{q^*}}} - \lambda \\
&\ge \sqrt{1 - (12/25)^2}\|\theta^{(i^*)}\|_2 \sqrt{\frac{n\tau}{6}} - \lambda > 0.358\sqrt{n\tau}\|\theta^{(i^*)}\|_2 - \lambda > \left(\frac{71.6}{\rho\tau} - 1\right)\lambda > \xi. \qquad (4.18)
\end{aligned}$$

Thus (ii) is satisfied with a new changepoint $b := s_{q_0} + \hat{z}^{[q_0]}$ detected. It remains to check that (i) holds for the pairs of arguments (s, b) and (b, e), for which it suffices to show that $\min_{1 \le i \le \nu} |b - z_i| \le n\rho$. To this end, we study the behaviour of univariate CUSUM statistics of the projected series $(\hat{v}^{[q_0]})^\top X^{[q_0]}$. To simplify notation, we define $\bar{X} := (\hat{v}^{[q_0]})^\top X^{[q_0]}$, $\bar{\mu} := (\hat{v}^{[q_0]})^\top \boldsymbol{\mu}^{[q_0]}$, $\bar{W} := (\hat{v}^{[q_0]})^\top W^{[q_0]}$, $\bar{T} := \mathcal{T}(\bar{X})$, $\bar{A} := \mathcal{T}(\bar{\mu})$ and $\bar{E} := \mathcal{T}(\bar{W})$. The row vector $\bar{\mu} \in \mathbb{R}^{e_{q_0} - s_{q_0}}$ is piecewise constant with changepoints at $z_{i_1+1} - s_{q_0}, \ldots, z_{i_2-1} - s_{q_0}$. Recall that $\hat{z}^{[q_0]} \in \operatorname{argmax}_{1 \le t \le e_{q_0} - s_{q_0} - 1} |\bar{T}_t|$. We may assume that $\bar{T}_{\hat{z}^{[q_0]}} > 0$ (the case $\bar{T}_{\hat{z}^{[q_0]}} < 0$ can be handled similarly). On Ω_3,

$$\bar{A}_{\hat{z}^{[q_0]}} = \bar{T}_{\hat{z}^{[q_0]}} - \bar{E}_{\hat{z}^{[q_0]}} \ge \bar{T}^{[q_0]}_{\max} - \lambda \ge \left(\frac{71.6}{\rho\tau} - 2\right)\lambda > 0, \qquad (4.19)$$

and in particular, there is at least one changepoint in $(s_{q_0}, e_{q_0}]$. We may assume that $\hat{z}^{[q_0]}$ is not equal to $z_i - s_{q_0}$ for any $i_1 + 1 \leq i \leq i_2 - 1$, since otherwise $\min_{1\leq i\leq \nu} |b - z_i| = 0$ and we are done. By Lemma 4.12 and after possibly reflecting the time direction, we may also assume that there is at least one changepoint to the left of $\hat{z}^{[q_0]}$, and that if $z_{i_0} - s_{q_0}$ is the changepoint immediately left of $\hat{z}^{[q_0]}$, then the series $\{\bar{A}_t : z_{i_0} - s_{q_0} \leq t \leq \hat{z}^{[q_0]}\}$ is positive and strictly decreasing. It therefore follows by (4.19), and (4.18) with i_0 in place of i^*, that on Ω_3,

$$\bar{A}_{z_{i_0}-s_{q_0}} \geq \bar{A}_{\hat{z}^{[q_0]}} \geq \bar{T}^{[q_0]}_{\max} - \lambda \geq 0.358\sqrt{n\tau}\|\theta^{(i_0)}\|_2 - 2\lambda > \left(0.358 - \frac{\rho\tau}{100}\right)\sqrt{n\tau}\|\theta^{(i_0)}\|_2 > \frac{71.4\lambda}{\rho\tau}, \tag{4.20}$$

where we use $9\rho < \tau \leq 1/2$ in the final inequality. On the other hand, on the event Ω_3, by the maximality of $\bar{T}_{\hat{z}^{[q_0]}}$, we have that

$$\bar{A}_{\hat{z}^{[q_0]}} \geq \bar{T}_{\hat{z}^{[q_0]}} - \lambda \geq \bar{T}_{z_{i_0}-s_{q_0}} - \lambda \geq \bar{A}_{z_{i_0}-s_{q_0}} - 2\lambda \geq \left(1 - \frac{\rho\tau}{35.7}\right)\bar{A}_{z_{i_0}-s_{q_0}}. \tag{4.21}$$

Our strategy here is to characterise the rate of decay of the series $\{\bar{A}_t : z_{i_0} - s_{q_0} \leq t \leq \hat{z}^{[q_0]}\}$ from its left endpoint, so that we can conclude from (4.21) that $\hat{z}^{[q_0]}$ is close to $z_{i_0} - s_{q_0}$. This is achieved by considering the following three cases: (a) there is no changepoint to the right of $\hat{z}^{[q_0]}$, i.e. $z_{i_0+1} \geq e_{q_0}$; (b) $z_{i_0+1} \leq e_{q_0} - 1$ and $\bar{A}_{z_{i_0}} \geq \bar{A}_{z_{i_0+1}}$; (c) $z_{i_0+1} \leq e_{q_0} - 1$ and $\bar{A}_{z_{i_0}} < \bar{A}_{z_{i_0+1}}$.

In case (a), we apply Lemma 4.11 with $e_{q_0} - s_{q_0}$ and $z_{i_0} - s_{q_0}$ taking the roles of n and z in the lemma respectively, while noting that

$$\left|(\hat{v}^{[q_0]})^\top \theta^{(i_0)}\right|^{-1} \bar{A}_t = \sqrt{\frac{e_{q_0} - s_{q_0} - t}{(e_{q_0} - s_{q_0})t}}(z_{i_0} - s_{q_0}) \qquad \forall\, z_{i_0} - s_{q_0} \leq t \leq e_{q_0} - s_{q_0}$$

takes the same form as the function f in the corresponding range in Lemma 4.11. Thus, we conclude from (4.21) and Lemma 4.11 that

$$\hat{z}^{[q_0]} - (z_{i_0} - s_{q_0}) \leq 2(e_{q_0} - z_{i_0})\frac{\rho\tau}{35.7} \leq n\rho,$$

as desired.

For case (b), define

$$\tilde{\mu} := \frac{1}{e_{q_0} - s_{q_0}} \sum_{t=1}^{e_{q_0}-s_{q_0}} \bar{\mu}_t$$

to be the overall average of the $\bar{\mu}$ series, and let

$$\tilde{\mu}_{\mathrm{L}} := \frac{1}{z_{i_0} - s_{q_0}} \sum_{t=1}^{z_{i_0}-s_{q_0}} \bar{\mu}_t - \tilde{\mu}, \quad \tilde{\mu}_{\mathrm{M}} := \bar{\mu}_{z_{i_0}+1-s_{q_0}} - \tilde{\mu} \text{ and } \tilde{\mu}_{\mathrm{R}} := \frac{1}{e_{q_0} - z_{i_0+1}} \sum_{t=z_{i_0+1}-s_{q_0}+1}^{e_{q_0}-s_{q_0}} \bar{\mu}_t - \tilde{\mu}$$

be the centred averages of the $\bar{\mu}$ series on the segments $(0, z_{i_0} - s_{q_0}]$, $(z_{i_0} - s_{q_0}, z_{i_0+1} - s_{q_0}]$ and $(z_{i_0+1} - s_{q_0}, e_{q_0} - s_{q_0}]$ respectively. Using (4.3), we have that for $z_{i_0} - s_{q_0} \leq t \leq z_{i_0+1} - s_{q_0}$,

$$\bar{A}_t = [\mathcal{T}(\bar{\mu})]_t = \sqrt{\frac{e_{q_0} - s_{q_0}}{t(e_{q_0} - s_{q_0} - t)}}\Big\{(z_{i_0} - s_{q_0})(-\tilde{\mu}_{\mathrm{L}}) + (t - z_{i_0} + s_{q_0})(-\tilde{\mu}_{\mathrm{M}})\Big\}\Big\}. \tag{4.22}$$

We claim that $z_{i_0} - s_{q_0} \geq n\tau/15$. For, if not, then in particular, $z_{i_0-1} < s_{q_0}$ and $\tilde{\mu}_{\mathrm{L}} = \bar{\mu}_{z_{i_0}-s_{q_0}} - \tilde{\mu}$.

By (4.22) and the fact that $\bar{A}_{z_{i_0}-s_{q_0}} > 0$, we have $\tilde{\mu}_{\mathrm{L}} < 0$. It follows from (4.20) that

$$\begin{aligned}
0.357\sqrt{n\tau}(\tilde{\mu}_{\mathrm{M}} - \tilde{\mu}_{\mathrm{L}}) &\leq \left(0.358 - \frac{\rho\tau}{100}\right)\sqrt{n\tau}|(\hat{v}^{[q_0]})^\top \theta^{(i_0)}| \leq \bar{A}_{z_{i_0}-s_{q_0}} \\
&= \sqrt{\frac{(e_{q_0} - s_{q_0})(z_{i_0} - s_{q_0})}{e_{q_0} - z_{i_0}}}(-\tilde{\mu}_{\mathrm{L}}) \\
&\leq \sqrt{\frac{n\tau + z_{i_0} - s_{q_0}}{n\tau}}\sqrt{z_{i_0} - s_{q_0}}(-\tilde{\mu}_{\mathrm{L}}) \leq \frac{4\sqrt{n\tau}}{15}(-\tilde{\mu}_{\mathrm{L}}),
\end{aligned}$$

which can be rearranged to give $-\tilde{\mu}_{\mathrm{M}} > 0.25(-\tilde{\mu}_{\mathrm{L}})$. Consequently,

$$\begin{aligned}
\bar{A}_{z_{i_0+1}-s_{q_0}} &= \sqrt{\frac{e_{q_0} - s_{q_0}}{(z_{i_0+1} - s_{q_0})(e_{q_0} - z_{i_0+1})}}\Big\{(-\tilde{\mu}_{\mathrm{L}})(z_{i_0} - s_{q_0}) + (-\tilde{\mu}_{\mathrm{M}})(z_{i_0+1} - z_{i_0})\Big\} \\
&> \sqrt{\frac{e_{q_0} - s_{q_0}}{(z_{i_0+1} - s_{q_0})(e_{q_0} - z_{i_0+1})}}\Big\{(-\tilde{\mu}_{\mathrm{L}})(z_{i_0} - s_{q_0}) + 0.25(-\tilde{\mu}_{\mathrm{L}})(z_{i_0+1} - z_{i_0})\Big\} \\
&\geq 0.25\sqrt{\frac{(e_{q_0} - s_{q_0})(z_{i_0+1} - s_{q_0})}{e_{q_0} - z_{i_0+1}}}(-\tilde{\mu}_{\mathrm{L}}) \geq 0.25\bar{A}_{z_{i_0}-s_{q_0}}\sqrt{\frac{z_{i_0+1} - s_{q_0}}{z_{i_0} - s_{q_0}}} \geq \bar{A}_{z_{i_0}-s_{q_0}},
\end{aligned}$$

contradicting the assumption of case (b). Hence we have established the claim. We can then apply Lemma 4.13 with $\bar{A}_t$, $e_{q_0} - s_{q_0}$, $z_{i_0} - s_{q_0}$, $z_{i_0+1} - s_{q_0}$, $-\tilde{\mu}_{\mathrm{L}}$, $-\tilde{\mu}_{\mathrm{M}}$ and $\tau/15$ taking the roles of $g(t)$, n, z, z', μ_0, μ_1 and τ in the lemma respectively. By (4.21) and this lemma, we conclude that

$$\hat{z}^{[q_0]} - (z_{i_0} - s_{q_0}) \leq \frac{\rho\tau\bar{A}_{z_{i_0}-s_{q_0}}/35.7}{0.52\bar{A}_{z_{i_0}-s_{q_0}}n^{-1}\tau/15} \leq n\rho.$$

For case (c), by Lemma 4.12, the series $(\bar{A}_t : z_{i_0} - s_{q_0} \leq t \leq z_{i_0+1} - s_{q_0})$ must be strictly decreasing, then strictly increasing, while staying positive throughout. Define $\zeta := \max\{t \in [z_{i_0} - s_{q_0}, z_{i_0+1} - s_{q_0}] : \bar{A}_t \leq \bar{A}_{z_{i_0+1}-s_{q_0}} - 2\lambda\}$. Using a very similar argument to that in case (b), we find that $e_{q_0} - z_{i_0+1} \geq n\tau/15$, and therefore by Lemma 4.13 again, $z_{i_0+1} - s_{q_0} - (\zeta + 1) \leq n\rho$. Now on Ω_3, we have $\bar{A}_{z_{i_0}-s_{q_0}} > \bar{A}_{\hat{z}^{[q_0]}} > \bar{A}_{z_{i_0+1}-s_{q_0}} - 2\lambda \geq \bar{A}_\zeta$ and $\zeta - (z_{i_0} - s_{q_0}) \geq n\tau - n\rho - 1$. So we can apply the same argument as in case (b) with ζ taking the role of z_{i_0+1} and $\tau - \rho - 1/n$ in place of τ, and obtain that

$$\hat{z}^{[q_0]} - (z_{i_0} - s_{q_0}) \leq \frac{\rho\tau\bar{A}_{z_{i_0}-s_{q_0}}/35.7}{0.52\bar{A}_{z_{i_0}-s_{q_0}}n^{-1}(\tau - \rho - 1/n)/15} \leq \frac{15/0.52}{35.7(1 - 1/9 - 1/14)}n\rho \leq n\rho,$$

as desired. □

4.7 Appendix: Ancillary results

The following result is a generalisation of the curvature lemma of Vu et al. (2013, Lemma 3.1).

Lemma 4.4. *Let $v \in \mathbb{S}^{p-1}$ and $u \in \mathbb{S}^{n-1}$ be the leading left and right singular vectors of $A \in \mathbb{R}^{p\times n}$ respectively. Suppose that the first and second largest singular values of A are separated by $\delta > 0$. Let $M \in \mathbb{R}^{p\times n}$. If either of the following two conditions holds,*

(a) $\mathrm{rank}(A) = 1$ *and* $\|M\|_2 \leq 1$,
(b) $\|M\|_* \leq 1$,

then

$$\|vu^\top - M\|_2^2 \leq \frac{2}{\delta}\langle A, vu^\top - M\rangle.$$

Remark: We note that if $v \in \mathbb{S}^{p-1}$ and $u \in \mathbb{S}^{n-1}$ are the leading left and right singular vectors respectively of $A \in \mathbb{R}^{p\times n}$, then since the matrix operator norm and the nuclear norm are dual norms with respect to the trace inner product, we have that $\langle A, vu^\top\rangle = v^\top Au = \|A\|_{\mathrm{op}} = \sup_{M\in\mathcal{S}_1}\langle A, M\rangle$. Thus, Lemma 4.4 provides a lower bound on the curvature of the function $M \mapsto \langle A, M\rangle$ as M moves away from the maximiser of the function in $\mathcal{S}_1$.

Proof. Let $A = VDU^\top$ be the singular value decomposition of A, where $V \in \mathbb{R}^{p\times p}$ and $U \in \mathbb{R}^{n\times n}$ are orthogonal matrices with column vectors $v_1 = v, v_2, \ldots, v_p$ and $u_1 = u, u_2, \ldots, u_n$ respectively, and $D \in \mathbb{R}^{p\times n}$ is a rectangular diagonal matrix with nonnegative entries along its main diagonal. The diagonal entries $\sigma_i := D_{ii}$ are the singular values of A, and we may assume without loss of generality that $\sigma_1 \geq \cdots \geq \sigma_r > 0$ are all the positive singular values, for some $r \leq \min\{n, p\}$.

Let $\tilde{M} := V^\top MU$ and denote $e_1^{[d]} := (1, 0, \ldots, 0)^\top \in \mathbb{R}^d$. Then by unitary invariance of the Frobenius norm, we have

$$\|v_1u_1^\top - M\|_2^2 = \|e_1^{[p]}(e_1^{[n]})^\top - \tilde{M}\|_2^2 = \|\tilde{M}\|_2^2 + 1 - 2\tilde{M}_{11}. \tag{4.23}$$

On the other hand,

$$\langle A, v_1u_1^\top - M\rangle = \langle D, e_1^{[p]}(e_1^{[n]})^\top - \tilde{M}\rangle = \sigma_1 - \sum_{i=1}^r \sigma_i\tilde{M}_{ii} \geq \sigma_1(1 - \tilde{M}_{11}) - \sigma_2\sum_{i=2}^r |\tilde{M}_{ii}|. \tag{4.24}$$

If condition (a) holds, then $\sigma_2 = 0$ and $\delta = \sigma_1$, so by (4.23) and (4.24), we have

$$\|v_1u_1^\top - M\|_2^2 \leq 2(1 - \tilde{M}_{11}) = \frac{2}{\delta}\langle A, v_1u_1^\top - M\rangle,$$

as desired.

On the other hand, if condition (b) holds, then by the characterisation of the nuclear norm in Lemma 4.7 as well as its unitary invariance, we have

$$\sum_{i=1}^r |\tilde{M}_{ii}| = \sup_{\substack{U\in\mathbb{R}^{p\times n}\text{ diagonal}\\ U_{ii}\in\{\pm1\}\ \forall i}} \langle U, \tilde{M}\rangle \leq \|\tilde{M}\|_* = \|M\|_* \leq 1. \tag{4.25}$$

On the other hand, if $\|M\|_* \leq 1$, then $\sigma_i \leq 1$ for all i, so

$$\|M\|_2 = \left(\sum_{i=1}^r \sigma_i^2\right)^{1/2} \leq \left(\sum_{i=1}^r \sigma_i\right)^{1/2} \leq 1. \tag{4.26}$$

Using (4.23), (4.24), (4.25) and (4.26), we therefore have

$$\begin{aligned}\langle A, v_1u_1^\top - M\rangle &\geq \sigma_1(1 - \tilde{M}_{11}) - \sigma_2\sum_{i=2}^r |\tilde{M}_{ii}| \geq (\sigma_1 - \sigma_2)(1 - \tilde{M}_{11})\\ &\geq \frac{\delta}{2}(\|\tilde{M}\|_2^2 + 1 - 2\tilde{M}_{11}) = \frac{\delta}{2}\|v_1u_1^\top - M\|_2^2,\end{aligned}$$

as desired. □

Lemma 4.5. *Let $W \in \mathbb{R}^{p\times n}$ have independent $N(0,\sigma^2)$ entries and let $E := \mathcal{T}(W)$. Then for $u > 0$, we have*

$$\mathbb{P}\big(\|E\|_\infty \geq u\sigma\big) \leq \sqrt{\frac{2}{\pi}} p\lceil \log n\rceil (u + 2/u) e^{-u^2/2}.$$

Proof. Let B be a standard Brownian bridge on $[0,1]$. Then for every $j \in \{1,\dots,p\}$,

$$(E_{j,1},\dots,E_{j,(n-1)}) \stackrel{d}{=} \left(\frac{\sigma B(t)}{\sqrt{t(1-t)}}\right)_{t=\frac{1}{n},\dots,\frac{n-1}{n}}.$$

Thus,

$$\mathbb{P}\big\{\|E\|_\infty \geq \sigma u\big\} \leq p\mathbb{P}\left\{\sup_{t\in[1/n,1-1/n]} \frac{B(t)}{\sqrt{t(1-t)}} \geq u\right\}.$$

Let $t = t(s) := e^{2s}/(e^{2s}+1)$ and define the process X by $X(s) := \{t(s)(1-t(s))\}^{-1/2} B(t(s))$. Recall that the Ornstein–Uhlenbeck process is the centred continuous Gaussian process $\{U(s) : s \in \mathbb{R}\}$ having covariance function $\mathrm{Cov}(U(s_1),U(s_2)) = e^{-|s_1-s_2|}$. We compute that

$$\begin{aligned}\mathrm{Cov}\big(X(s_1),X(s_2)\big) &= \mathrm{Cov}\left(\frac{B\big(e^{2s_1}/(e^{2s_1}+1)\big)}{\sqrt{e^{2s_1}/(e^{2s_1}+1)^2}}, \frac{B\big(e^{2s_2}/(e^{2s_2}+1)\big)}{\sqrt{e^{2s_2}/(e^{2s_2}+1)^2}}\right)\\ &= \left(\frac{e^{s_1}}{e^{2s_1}+1}\frac{e^{s_2}}{e^{2s_2}+1}\right)^{-1} \frac{e^{2\min(s_1,s_2)}}{e^{2\min(s_1,s_2)}+1}\frac{1}{e^{2\max(s_1,s_2)}+1} = e^{-|s_1-s_2|}.\end{aligned}$$

Thus, X is the Ornstein–Uhlenbeck process and we have

$$\mathbb{P}\left\{\sup_{t\in[1/n,1-1/n]}\left|\frac{B(t)}{\sqrt{t(1-t)}}\right| \geq u\right\} = \mathbb{P}\left\{\sup_{s\in[0,\log(n-1)]}|X(s)| \geq u\right\} \leq \lceil \log n\rceil \mathbb{P}\left\{\sup_{s\in[0,1]}|X(s)| \geq u\right\},$$

where the inequality follows from the stationarity of the Ornstein–Uhlenbeck process and a union bound. Let $Y = \{Y(t) : t \in \mathbb{R}\}$ be a centred continuous Gaussian process with covariance function $\mathrm{Cov}(Y(s),Y(t)) = \max(1-|s-t|,0)$. Since $\mathbb{E}X(t)^2 = \mathbb{E}Y(t)^2 = 1$ for all t and $\mathrm{Cov}(X(s),X(t)) \geq \mathrm{Cov}(Y(s),Y(t))$, by Slepian's inequality (Slepian, 1962), $\sup_{s\in[0,1]}|Y(s)|$ stochastically dominates $\sup_{s\in[0,1]}|X(s)|$. Hence it suffices to establish the required bound with Y in place of X. The process Y, known as the Slepian process, has excursion probabilities given by closed-form expressions (Slepian, 1961; Shepp, 1971): for $x < u$,

$$\mathbb{P}\left\{\sup_{s\in[0,1]} Y(s) \geq u \;\middle|\; Y(0) = x\right\} = 1 - \Phi(u) + \frac{\phi(u)}{\phi(x)}\Phi(x),$$

where ϕ and Φ are respectively the density and distribution functions of the standard normal distribution. Hence for $u > 0$ we can write

$$\begin{aligned}\mathbb{P}\left\{\sup_{s\in[0,1]}|Y(s)| \geq u\right\} &= \int_{-\infty}^{\infty} \mathbb{P}\left\{\sup_{s\in[0,1]}|Y(s)| \geq u \;\middle|\; Y(0) = x\right\}\phi(x)\,dx\\ &\leq \mathbb{P}(|Y(0)| \geq u) + 2\int_{-u}^{u} \mathbb{P}\left\{\sup_{s\in[0,1]} Y(s) \geq u \;\middle|\; Y(0) = x\right\}\phi(x)\,dx\\ &= 2\Phi(-u) + 2\int_{-u}^{u}\big\{\phi(x)\Phi(-u) + \phi(u)\Phi(x)\big\}\,dx\\ &= 2u\phi(u) + 4\Phi(-u)\{1-\Phi(-u)\} \leq 2(u + 2u^{-1})\phi(u),\end{aligned}$$

as desired. □

Proposition 4.6. *Suppose the first and second largest singular values of $A \in \mathbb{R}^{p\times n}$ are separated by $\delta > 0$. Let unit vectors $v \in \mathbb{S}^{p-1}(k)$ and $u \in \mathbb{S}^{n-1}(\ell)$ be left and right leading singular vectors of A respectively. Let $T \in \mathbb{R}^{p\times n}$ satisfy $\|T - A\|_\infty \le \lambda$ for some $\lambda > 0$, and let $\mathcal{S}$ be a subset of $p \times n$ real matrices containing $vu^\top$. Suppose one of the following two conditions holds:*

(a) $\operatorname{rank}(A) = 1$ *and* $\mathcal{S} \subseteq \{M \in \mathbb{R}^{p\times n} : \|M\|_2 \le 1\}$
(b) $\mathcal{S} \subseteq \{M \in \mathbb{R}^{p\times n} : \|M\|_* \le 1\}$.

Then for any

$$\hat{M} \in \operatorname*{argmax}_{M\in\mathcal{S}} \big\{ \langle T, M\rangle - \lambda\|M\|_1 \big\},$$

we have

$$\|vu^\top - \hat{M}\|_2 \le \frac{4\lambda\sqrt{k\ell}}{\delta}.$$

Furthermore, if $\hat{v}$ and $\hat{u}$ are leading left and right singular vectors of $\hat{M}$ respectively, then

$$\max\{\sin\angle(\hat{v}, v), \sin\angle(\hat{u}, u)\} \le \frac{8\lambda\sqrt{k\ell}}{\delta}. \tag{4.27}$$

Proof. Using Lemma 4.4, we have

$$\|vu^\top - \hat{M}\|_2^2 \le \frac{2}{\delta}\langle A, vu^\top - \hat{M}\rangle = \frac{2}{\delta}\big(\langle T, vu^\top - \hat{M}\rangle + \langle A - T, vu^\top - \hat{M}\rangle\big). \tag{4.28}$$

Since $\hat{M}$ is a maximiser of the objective function $M \mapsto \langle T, M\rangle - \lambda\|M\|_1$ over the set $\mathcal{S}$, and since $vu^\top \in \mathcal{S}$, we have the basic inequality

$$\langle T, vu^\top - \hat{M}\rangle \le \lambda(\|vu^\top\|_1 - \|\hat{M}\|_1). \tag{4.29}$$

Denote $S_v := \{j : 1 \le j \le p, v_j \ne 0\}$ and $S_u := \{t : 1 \le t \le n, u_t \ne 0\}$. From (4.28) and (4.29) and the fact that $\|T - A\|_\infty \le \lambda$, we have

$$\begin{aligned}
\|vu^\top - \hat{M}\|_2^2 &\le \frac{2}{\delta}\big(\lambda\|vu^\top\|_1 - \lambda\|\hat{M}\|_1 + \lambda\|vu^\top - \hat{M}\|_1\big) \\
&= \frac{2\lambda}{\delta}\big(\|v_{S_v}u_{S_u}^\top\|_1 - \|\hat{M}_{S_vS_u}\|_1 + \|v_{S_v}u_{S_u}^\top - \hat{M}_{S_vS_u}\|_1\big) \\
&\le \frac{4\lambda}{\delta}\|v_{S_v}u_{S_u}^\top - \hat{M}_{S_vS_u}\|_1 \le \frac{4\lambda\sqrt{k\ell}}{\delta}\|vu^\top - \hat{M}\|_2.
\end{aligned}$$

Dividing through by $\|vu^\top - \hat{M}\|_2$, we have the first desired result.

Now, by definition of the operator norm, we have

$$\begin{aligned}
\|vu^\top - \hat{M}\|_2^2 &= 1 + \|\hat{M}\|_2^2 - 2v^\top\hat{M}u \\
&\ge 1 + \|\hat{M}\|_2^2 - 2\|\hat{M}\|_{\mathrm{op}} = 1 + \|\hat{M}\|_2^2 - 2\hat{v}^\top\hat{M}\hat{u} = \|\hat{v}\hat{u}^\top - \hat{M}\|_2^2.
\end{aligned}$$

Thus,

$$\|vu^\top - \hat{v}\hat{u}^\top\|_2 \le \|vu^\top - \hat{M}\|_2 + \|\hat{v}\hat{u}^\top - \hat{M}\|_2 \le 2\|vu^\top - \hat{M}\|_2 \le \frac{8\lambda\sqrt{k\ell}}{\delta}. \tag{4.30}$$

We claim that

$$\max\big\{\sin^2\angle(\hat{u}, u), \sin^2\angle(\hat{v}, v)\big\} \le \|vu^\top - \hat{v}\hat{u}^\top\|_2^2. \tag{4.31}$$

Let $v_0 := (v + \hat{v})/2$ and $\Delta := v - v_0$. Then

$$\begin{aligned}\|vu^\top - \hat{v}\hat{u}^\top\|_2^2 &= \|(v_0 + \Delta)u^\top - (v_0 - \Delta)\hat{u}^\top\|_2^2 = \|v_0(u - \hat{u})^\top\|_2^2 + \|\Delta(u + \hat{u})^\top\|_2^2 \\ &= \|v_0\|_2^2\|u - \hat{u}\|_2^2 + \|\Delta\|_2^2\|u + \hat{u}\|_2^2 \\ &\geq (\|v_0\|_2^2 + \|\Delta\|_2^2)\min(\|u - \hat{u}\|_2^2, \|u + \hat{u}\|_2^2) \\ &\geq \left\{1 - (\hat{u}^\top u)^2\right\} = \sin^2 \angle(\hat{u}, u),\end{aligned}$$

where the penultimate step uses the fact that $\|v_0\|_2^2 + \|\Delta\|_2^2 = 1$. A similar inequality holds for $\sin^2 \angle(\hat{v}, v)$, which establishes the desired claim (4.31). Inequality (4.27) now follows from (4.30) and (4.31). □

The lemma below gives a characterisation of the nuclear norm of a real matrix.

Lemma 4.7. *For $n, p \geq 1$, let $\mathcal{V}_n$ and $\mathcal{V}_p$ be respectively the sets of $n \times \min(n, p)$ and $p \times \min(n, p)$ real matrices having orthonormal columns. Let $A \in \mathbb{R}^{p \times n}$. Then*

$$\|A\|_* = \sup_{V \in \mathcal{V}_p, U \in \mathcal{V}_n} \langle VU^\top, A \rangle$$

Proof. Suppose we have the singular value decomposition $A = \tilde{V}D\tilde{U}^\top$ where $\tilde{V} \in \mathcal{V}_p$, $\tilde{U} \in \mathcal{V}_n$ and where $D = (D_{ij}) \in \mathbb{R}^{\min(n,p) \times \min(n,p)}$ is a diagonal matrix with decreasing non-negative diagonal entries. Write V_j for the jth column of V and similarly U_j for the jth column of U. Then

$$\begin{aligned}\sup_{V \in \mathcal{V}_p, U \in \mathcal{V}_n} \langle VU^\top, A \rangle &= \sup_{V \in \mathcal{V}_p, U \in \mathcal{V}_n} \langle VU^\top, \tilde{V}D\tilde{U}^\top \rangle = \sup_{V \in \mathcal{V}_p, U \in \mathcal{V}_n} \langle VU^\top, D \rangle \\ &= \sup_{V \in \mathcal{V}_p, U \in \mathcal{V}_n} \sum_{j=1}^{\min(n,p)} D_{jj} V_j^\top U_j = \sum_{j=1}^{\min(n,p)} D_{jj} = \|A\|_*,\end{aligned}$$

as desired. □

Lemma 4.8. *Let $\gamma \in \mathbb{R}^{n-1}$ be defined as in (4.5) for some $n \geq 6$ and $2 \leq z \leq n - 2$. Let $\tilde{z} := \min(z, n - z)$. Then*

$$\begin{aligned}\frac{1}{4}\tilde{z} \leq \|\gamma\|_2 &\leq \sqrt{\log(en/2)}\,\tilde{z} \\ \frac{1}{2}\sqrt{n}\tilde{z} \leq \|\gamma\|_1 &\leq 2.1\sqrt{n}\tilde{z}.\end{aligned}$$

Proof. Since the norms of γ are invariant under substitution $z \mapsto n - z$, we may assume without loss of generality that $z \leq n - z$. Hence $\tilde{z} = z$. We have that

$$\begin{aligned}\|\gamma\|_2^2 &= \frac{1}{n}\left\{\sum_{t=1}^{z} \frac{t(n-z)^2}{n-t} + \sum_{t=z+1}^{n-1} \frac{(n-t)z^2}{t}\right\} \\ &= n^2\left\{\sum_{t=1}^{z} \frac{(t/n)(1-z/n)^2}{(1-t/n)} \cdot \frac{1}{n} + \sum_{t=z+1}^{n-1} \frac{(1-t/n)(z/n)^2}{t/n} \cdot \frac{1}{n}\right\},\end{aligned}$$

where the expression inside the bracket can be interpreted as a Riemann sum approximation to an

integral. We therefore find that

$$n^2\Big\{I_1 - \frac{(z/n)(1-z/n)}{n}\Big\} \le \|\gamma\|_2^2 \le n^2\Big\{I_1 + \frac{(z/n)(1-z/n)}{n}\Big\},$$

where

$$\begin{aligned} I_1 &:= (1-z/n)^2 \int_0^{z/n} \frac{r}{1-r}\,dr + (z/n)^2 \int_{z/n}^1 \frac{1-r}{r}\,dr \\ &= (1-z/n)^2\big\{-\log(1-z/n) - z/n\big\} + (z/n)^2\big\{-\log(z/n) - (1-z/n)\big\}. \end{aligned}$$

Since $-\log(1-x) \ge x + x^2/2$ for $0 \le x < 1$, we have $I_1 \ge (z/n)^2(1-z/n)^2$. When $n \ge 6$ and $2 \le z \le n/2$, we find $\frac{(z/n)(1-z/n)}{n} \le 3I_1/4$. Hence,

$$\|\gamma\|_2 \ge \frac{1}{2} n (z/n)(1-z/n) \ge \frac{1}{4} z.$$

On the other hand, under the assumption that $z \le n/2$, we have $-\log(1-z/n) - z/n \le (z/n)^2$. Hence

$$\|\gamma\|_2^2 \le n^2\big\{(1-z/n)^2(z/n)^2 + (z/n)^2 \log(n/2)\big\} \le z^2 \log(en/2),$$

as required.

For the ℓ_1 norm, we similarly write $\|\gamma\|_1$ as a Riemann sum:

$$\begin{aligned} \|\gamma\|_1 &= \frac{1}{\sqrt{n}}\Big\{\sum_{t=1}^{z} \sqrt{\frac{t}{n-t}}(n-z) + \sum_{t=z+1}^{n-1} \sqrt{\frac{(n-t)}{t}} z\Big\} \\ &= n^{3/2}\Big\{\sum_{t=1}^{z} \sqrt{\frac{t/n}{1-t/n}}(1-z/n)\cdot\frac{1}{n} + \sum_{t=z+1}^{n-1} \frac{1-t/n}{t/n}(z/n)\cdot\frac{1}{n}\Big\}. \end{aligned}$$

So

$$n^{3/2}\Big\{I_2 - \frac{\sqrt{z/n(1-z/n)}}{n}\Big\} \le \|\gamma\|_1 \le n^{3/2}\Big\{I_2 + \frac{\sqrt{z/n(1-z/n)}}{n}\Big\},$$

where

$$I_2 := (1-z/n)\int_0^{z/n} \sqrt{\frac{r}{1-r}}\,dr + (z/n)\int_{z/n}^1 \sqrt{\frac{1-r}{r}}\,dr = (1-z/n)g(z/n) + (z/n)g(1-z/n),$$

where function $g(a) := \int_0^a \sqrt{r/(1-r)}\,dr = \arcsin(\sqrt{a}) - \sqrt{a(1-a)}$. We can check that $g(a)/a^{3/2}$ has positive first derivative throughout $(0,1)$, and $g(a)/a^{3/2} \searrow 2/3$ as $a \searrow 0$. This implies that $2a^{3/2}/3 \le g(a) \le \pi a^{3/2}/2$. Consequently,

$$\frac{2z}{3n}\Big(1-\frac{z}{n}\Big)\Big(\sqrt{\frac{z}{n}} + \sqrt{1-\frac{z}{n}}\Big) \le I_2 \le \frac{\pi}{2}\frac{z}{n}\Big(1-\frac{z}{n}\Big)\Big(\sqrt{\frac{z}{n}} + \sqrt{1-\frac{z}{n}}\Big)$$

Also, for $n \ge 6$ and $2 \le z \le n/2$,

$$\frac{\sqrt{z/n(1-z/n)}}{n} \le \frac{\sqrt{3}}{4+2\sqrt{2}}\frac{z}{n}\Big(1-\frac{z}{n}\Big)\Big(\sqrt{\frac{z}{n}} + \sqrt{1-\frac{z}{n}}\Big).$$

Therefore,

$$\|\gamma\|_1 \leq (\pi/2 + \sqrt{3}/(4+2\sqrt{2}))\sqrt{n}z \sup_{0\leq y\leq 1/2}(1-y)(\sqrt{y}+\sqrt{1-y}) \leq 2.1\sqrt{n}z,$$

and

$$\|\gamma\|_1 \geq (1 - \sqrt{3}/(4+2\sqrt{2}))\sqrt{n}z \inf_{0\leq y\leq 1/2}(1-y)(\sqrt{y}+\sqrt{1-y}) \geq \frac{1}{2}\sqrt{n}z,$$

□

Proposition 4.9. *Let $X \sim P \in \mathcal{P}(n, p, k, 1, \vartheta, \tau, \sigma^2)$, with the single changepoint located at z, say (so we may take $\tau = n^{-1}\min\{z, n-z\}$). Define A, E and T as in Section 4.3. Let $v \in \operatorname{argmax}_{\tilde{v}\in\mathbb{S}^{p-1}} \|A^\top \tilde{v}\|_2$ and $\hat{v} \in \operatorname{argmax}_{\tilde{v}\in\mathbb{S}^{p-1}(k)} \|T^\top \tilde{v}\|_2$. If $n \geq 6$, then with probability at least $1 - 4(p\log n)^{-1/2}$,*

$$\sin\angle(\hat{v}, v) \leq \frac{16\sqrt{2}\sigma}{\tau\vartheta}\sqrt{\frac{\log(p\log n)}{n}}.$$

Proof. From the definition in Section 4.3, $A = \theta\gamma^\top$, for some $\theta \in \mathbb{R}^p$ satisfying $\|\theta\|_0 \leq k$ and $\|\theta\|_2^2 \geq k\vartheta^2$ and γ defined by (4.5). Then we have $v = \theta/\|\theta\|_2$. Define also $u := \gamma/\|\gamma\|_2$ and $\hat{u} := T^\top\hat{v}/\|T^\top\hat{v}\|_2$. Then by definition of $\hat{v}$, we have

$$\langle \hat{v}\hat{u}^\top, T\rangle = \|T^\top\hat{v}\|_2 \geq v^\top T u = \langle vu^\top, T\rangle. \tag{4.32}$$

By Lemma 4.4 and (4.32), we obtain

$$\begin{aligned}\|vu^\top - \hat{v}\hat{u}^\top\|_2^2 &\leq \frac{2}{\|\theta\|_2\|\gamma\|_2}\langle A, vu^\top - \hat{v}\hat{u}^\top\rangle \\ &\leq \frac{2}{\|\theta\|_2\|\gamma\|_2}\langle A - T, vu^\top - \hat{v}\hat{u}^\top\rangle \leq \frac{2}{\|\theta\|_2\|\gamma\|_2}\|E\|_\infty\|vu^\top - \hat{v}\hat{u}^\top\|_1. \end{aligned}\tag{4.33}$$

Note that in fact $v \in \mathbb{S}^{p-1}(k)$, by definition of the matrix A. Moreover, $\hat{v} \in \mathbb{S}^{p-1}(k)$ too, so the matrix $vu^\top - \hat{v}\hat{u}^\top$ has at most $2k$ non-zero rows. Thus, by the Cauchy–Schwarz inequality,

$$\|vu^\top - \hat{v}\hat{u}^\top\|_1 \leq \sqrt{2kn}\|vu^\top - \hat{v}\hat{u}^\top\|_2.$$

By (4.31) in the proof of Proposition 4.6, and (4.33), we find that

$$\sin\angle(\hat{v}, v) \leq \|vu^\top - \hat{v}\hat{u}^\top\|_2 \leq \frac{2\sqrt{2}\|E\|_\infty\sqrt{kn}}{\|\theta\|_2\|\gamma\|_2} \leq \frac{8\sqrt{2}\|E\|_\infty}{\vartheta\tau\sqrt{n}},$$

where we have used Lemma 4.8 in the final inequality. The desired result follows from bounding $\|E\|_\infty$ with high probability as in (4.14). □

Lemma 4.10. *Let $T \in \mathbb{R}^{p\times(n-1)}$ and $\lambda > 0$. Then the following optimisation problem*

$$\max_{M\in\mathcal{S}_2}\left\{\langle T, M\rangle - \lambda\|M\|_1\right\}$$

has a unique solution given by

$$\tilde{M} = \frac{\mathbf{soft}(T, \lambda)}{\|\mathbf{soft}(T, \lambda)\|_2}. \tag{4.34}$$

Proof. Define $\phi(M, R) := \langle T - R, M\rangle$ and $\mathcal{R} := \{R \in \mathbb{R}^{p\times(n-1)} : \|R\|_\infty \leq \lambda\}$. Then the objective function in the lemma is given by

$$f(M) = \min_{R\in\mathcal{R}} \phi(M, R).$$

We also define

$$g(R) := \max_{M\in\mathcal{S}_2} \phi(M, R) = \|T - R\|_2.$$

Since $\mathcal{S}_2$ and $\mathcal{R}$ are compact, convex subsets of $\mathbb{R}^{p\times(n-1)}$ endowed with the trace inner product, and since ϕ is affine and continuous in both M and R, we can use the minimax equality theorem Fan (1953, Theorem 1) to obtain

$$\max_{M\in\mathcal{S}_2} f(M) = \max_{M\in\mathcal{S}_2} \min_{R\in\mathcal{R}} \phi(M, R) = \min_{R\in\mathcal{R}} \max_{M\in\mathcal{S}_2} \phi(M, R) = \min_{R\in\mathcal{R}} g(R).$$

We note that the dual function g has a unique minimum over $\mathcal{R}$ at $R^{(d)}$, say, where $R^{(d)}_{j,t} := \mathrm{sgn}(T_{j,t}) \min(\lambda, |T_{j,t}|)$. Let

$$M^{(d)} \in \operatorname*{argmax}_{M\in\mathcal{S}_2} \phi(M, R^{(d)}), \quad M^{(p)} \in \operatorname*{argmax}_{M\in\mathcal{S}_2} f(M) \quad \text{and} \quad R^{(p)} \in \operatorname*{argmin}_{R\in\mathcal{R}} \phi(M^{(p)}, R).$$

Then

$$\min_{R\in\mathcal{R}} g(R) = \langle T - R^{(d)}, M^{(d)}\rangle \geq \langle T - R^{(d)}, M^{(p)}\rangle \geq \langle T - R^{(p)}, M^{(p)}\rangle = \max_{M\in\mathcal{S}_2} f(M).$$

Since the two extreme ends of the chain of inequalities are equal, we necessarily have

$$R^{(d)} \in \operatorname*{argmin}_{R\in\mathcal{R}} \langle T - R, M^{(p)}\rangle,$$

and consequently,

$$M^{(p)} \in \operatorname*{argmax}_{M\in\mathcal{S}_2} \langle T - R^{(d)}, M\rangle.$$

The objective $M \mapsto \langle T - R^{(d)}, M\rangle = \langle \mathbf{soft}(T, \lambda), M\rangle$ has a unique maximiser at $\tilde{M}$ defined in (4.34). Thus, $M^{(p)}$ is unique and has the form given in the lemma. □

The following lemma is used to control the rate of decay of the univariate CUSUM statistic from its peak in the single changepoint setting.

Lemma 4.11. *For $n \in \mathbb{N}$ and $z \in \{0, 1, \ldots, n\}$, define $f : [0, n] \to \mathbb{R}$ by*

$$f(t) := \begin{cases} \sqrt{\frac{t}{n(n-t)}}(n - z), & \text{if } t \leq z \\ \sqrt{\frac{n-t}{nt}}\, z, & \text{if } t > z. \end{cases}$$

Then for $\Delta \leq 1$,

$$\{t : f(t) \geq f(z)(1 - \Delta)\} \subseteq [z - 2z\Delta, z + 2(n - z)\Delta].$$

Proof. We note first that $f(t)$ is maximised at $t = z$. If $t \leq z$ and $f(t) \geq f(z)(1 - \Delta)$, then

$$\frac{t}{n - t} \geq \frac{z}{n - z}(1 - \Delta)^2.$$

Equivalently,

$$t \geq \frac{nz(1-\Delta)^2}{n - z + z(1-\Delta)^2} \geq z(1-\Delta)^2 \geq z - 2z\Delta. \tag{4.35}$$

By symmetry, for $t > z$ with $f(t) \geq f(z)(1-\Delta)$, we have that

$$n - t \geq (n-z)(1-\Delta)^2 \geq n - z - 2(n-z)\Delta. \tag{4.36}$$

Combining (4.35) and (4.36), we have the desired result. □

Lemma 4.12. *Suppose that $0 = z_0 < z_1 < \cdots < z_\nu < z_{\nu+1} = n$ are integers and that $\mu \in \mathbb{R}^n$ satisfies $\mu_t = \mu_{t'}$ for all $z_i < t \leq t' \leq z_{i+1}$, $0 \leq i \leq \nu$. Define $A := \mathcal{T}(\mu) \in \mathbb{R}^{n-1}$, where we treat μ as a row vector. If the series $(A_t : z_i + 1 \leq t \leq z_{i+1})$ is not constantly zero, then one of the following is true:*

(a) $i = 0$ and $(A_t : z_i + 1 \leq t \leq z_{i+1})$ does not change sign and has strictly increasing absolute values,
(b) $i = \nu$ and $(A_t : z_i + 1 \leq t \leq z_{i+1})$ does not change sign and has strictly decreasing absolute values,
(c) $1 \leq i \leq \nu - 1$ and $(A_t : z_i + 1 \leq t \leq z_{i+1})$ is strictly monotonic,
(d) $1 \leq i \leq \nu - 1$ and $(A_t : z_i + 1 \leq t \leq z_{i+1})$ does not change sign and its absolute values are strictly decreasing then strictly increasing.

Proof. This follows from the proof of Venkatraman (1992, Lemma 2.2). □

The following lemma is used to control the rate of decay of the univariate CUSUM statistic of the mean series away from its maximum absolute value in the case of two changepoints.

Lemma 4.13. *Let $1 \leq z < z' \leq n - 1$ be integers and $\mu_0, \mu_1 \in \mathbb{R}$. Define $g : [z, z'] \to \mathbb{R}$ by*

$$g(y) := \sqrt{\frac{n}{y(n-y)}}\{z\mu_0 + (y - z)\mu_1\}$$

Suppose that $\min\{z, z' - z\} \geq n\tau$ and

$$G := \max_{y \in [z, z']} |g(y)| = g(z). \tag{4.37}$$

Then

$$\sup_{y \in [z, z + 0.2n\tau]} g'(y) \leq -0.52Gn^{-1}\tau.$$

Proof. Define $r := z/n$, $r' := z'/n$, $B := r(\mu_0 - \mu_1)$ and $f(x) := n^{-1/2}g(nx)$ for $x \in [r, r']$. Then

$$f(x) = \frac{B + \mu_1 x}{\sqrt{x(1-x)}} \quad \text{and} \quad f'(x) = \frac{(\mu_1 + 2B)x - B}{2\{x(1-x)\}^{3/2}}.$$

Condition (4.37) is equivalent to

$$Gn^{-1/2} = \max_{x \in [r, r']} |f(x)| = f(r) = \frac{r\mu_0}{\sqrt{r(1-r)}}. \tag{4.38}$$

The desired result of the lemma is equivalent to

$$\sup_{x\in[r,r+0.2\tau]} f'(x) \le -0.52Gn^{-1/2}\tau.$$

We may assume without loss of generality that it is not the case that $\mu_0 = \mu_1 = 0$, because otherwise f is the zero function and $G = 0$, so the result holds. In that case, $G > 0$, so $\mu_0 > 0$, and we prove the above inequality by considering the following three cases.

Case 1: $B \le 0$. Then $\mu_1 \ge \mu_0$ and in fact $\mu_1 + 2B < 0$, because otherwise f' is non-negative on $[r, r']$, and if $f'(r) = 0$ (which is the only remaining possibility from (4.38)) then $B = 0$ and $\mu_1 = 0$, so $\mu_0 = 0$, a contradiction. Moreover, since $\mathrm{sgn}(f'(x)) = \mathrm{sgn}\big((\mu_1 + 2B)x - B\big)$, we deduce that $\frac{B}{\mu_1+2B} \le r \le 1$. In particular, $\mu_1 \le -B = r(\mu_1 - \mu_0) \le \mu_1 - \mu_0$ and hence $\mu_0 \le 0$, again a contradiction.

Case 2: $B > 0$ and $\mu_1 + 2B \le 0$. By (4.38) and the fact that $\mu_1 < 0$, so that $B > r\mu_0$, we have for $x \in [r, r + \tau]$ that

$$\begin{aligned} f'(x) &\le \frac{-B}{2\{x(1-x)\}^{3/2}} \\ &\le \frac{-B}{2\{r(1-r)\}^{1/2}} \inf_{x\in[r,r+\tau]} \frac{\{r(1-r)\}^{1/2}}{\{x(1-x)\}^{3/2}} \le -2Gn^{-1/2} \inf_{x\in[r,r+\tau]} \frac{r^{1/2}}{x^{1/2}} \le -\sqrt{2}Gn^{-1/2}. \end{aligned}$$

Here, we used the fact that $\min\{r, r' - r\} \ge \tau$ in the final bound.

Case 3: $B > 0$ and $\mu_1 + 2B > 0$, so that $\mu_0 > \mu_1$. In this case, considering $\mathrm{sgn}(f'(x))$ again yields $r \le \frac{B}{\mu_1+2B}$. We claim that

$$\frac{B}{\mu_1 + 2B} \ge r + 0.4\tau. \tag{4.39}$$

By the fundamental theorem of calculus,

$$\begin{aligned} f(r) - f\Big(\frac{B}{\mu_1+2B}\Big) &= \int_r^{\frac{B}{\mu_1+2B}} \frac{B - (\mu_1+2B)x}{2\{x(1-x)\}^{3/2}}\,dx \\ &= (\mu_1+2B)\Big(\frac{B}{\mu_1+2B} - r\Big)^2 \int_0^1 \frac{u}{2\{x(u)(1-x(u))\}^{3/2}}\,du, \end{aligned} \tag{4.40}$$

where we have used the substitution $x = x(u) := \frac{B}{\mu_1+2B} - (\frac{B}{\mu_1+2B} - r)u$ in the second step. Similarly,

$$\begin{aligned} f(r+\tau) - f\Big(\frac{B}{\mu_1+2B}\Big) &= \int_{\frac{B}{\mu_1+2B}}^{r+\tau} \frac{B - (\mu_1+2B)\tilde{x}}{2\{\tilde{x}(1-\tilde{x})\}^{3/2}}\,d\tilde{x} \\ &= (\mu_1+2B)\Big(r + \tau - \frac{B}{\mu_1+2B}\Big)^2 \int_0^1 \frac{u}{2\{\tilde{x}(u)(1-\tilde{x}(u))\}^{3/2}}\,du, \end{aligned} \tag{4.41}$$

using the substitution $\tilde{x} = \tilde{x}(u) := \frac{B}{\mu_1+2B} + (r + \tau - \frac{B}{\mu_1+2B})u$. For every $u \in [0, 1]$, we have $x(u) \le \tilde{x}(u) \le (1 + u)x(u)$. It follows that

$$\begin{aligned} \frac{\int_0^1 u\{\tilde{x}(u)(1-\tilde{x}(u))\}^{-3/2}\,du}{\int_0^1 u\{x(u)(1-x(u))\}^{-3/2}\,du} &\ge \frac{\int_0^1 ux(u)^{-3/2}(1+u)^{-3/2}\,du}{\int_0^1 ux(u)^{-3/2}\,du} = \frac{1}{2^{1/2}}\left\{\frac{(\frac{B}{\mu_1+2B})^{1/2} + r^{1/2}}{(\frac{2B}{\mu_1+2B})^{1/2} + r^{1/2}}\right\}^2 \\ &\ge \frac{1}{2^{1/2}}\left\{\frac{(r+\tau)^{1/2} + r^{1/2}}{2^{1/2}(r+\tau) + r^{1/2}}\right\}^2 \ge 0.45. \end{aligned} \tag{4.42}$$

Therefore, using (4.40), (4.41) and (4.42), together with the fact that $f(r) \geq f(r+\tau)$, we deduce that

$$\frac{B}{\mu_1 + 2B} - r \geq \frac{\tau}{1 + 0.45^{-1/2}} > 0.4\tau.$$

Hence (4.39) holds. For $x \in [r, r + 0.2\tau]$, we have

$$f'(x) \leq \frac{-(\mu_1 + 2B)\big(\frac{B}{2(\mu_1+2B)} - \frac{r}{2}\big)}{2\{x(1-x)\}^{3/2}} \leq \frac{-0.4\tau(\mu_1 + 2B)}{\sqrt{1.2r(1-r)}}. \tag{4.43}$$

If $\mu_1 \geq 0$, then $r \leq \frac{B}{\mu_1+2B} \leq 1/2$ and

$$\mu_1 + 2B = 2r\mu_0 + (1-2r)\mu_1 \geq 2r\mu_0. \tag{4.44}$$

If $\mu_1 < 0$ and $r \geq 1/2$, then

$$\mu_1 + 2B = 2r\mu_0 + (2r-1)(-\mu_1) \geq 2r\mu_0. \tag{4.45}$$

Finally, if $\mu_1 < 0$ and $r < 1/2$, then, writing $a := 1 - 2r$ and $b := \frac{2B}{\mu_1+2B} - 1$, we have from (4.39) that $a + b \geq 0.8\tau$ and

$$\begin{aligned}(\mu_1 + 2B)\left(\frac{B}{\mu_1 + 2B} - r\right) &= r(1-2r)\mu_0 - 2r(1-r)\mu_1 = ar\mu_0 + \frac{(1-a^2)B}{1+b^{-1}} \\ &\geq \left(a + \frac{1-a^2}{1 + (0.8\tau - a)^{-1}}\right) r\mu_0 \geq 0.57\tau r\mu_0. \end{aligned} \tag{4.46}$$

It follows from (4.43), (4.44), (4.45), (4.46) and (4.38) that for $x \in [r, r + 0.2\tau]$,

$$f'(x) \leq \frac{-0.57\tau r\mu_0}{\sqrt{1.2r(1-r)}} \leq -0.52Gn^{-1/2}\tau,$$

as desired. □

Bibliography

Alon, N., Andoni, A., Kaufman, T., Matulef, K., Rubinfeld, R., and Xie, N. (2007) Testing k-wise and almost k-wise independence. *Proceedings of the Thirty-Ninth ACM Symposium on Theory of Computing*, 496–505.

Alon, N., Krivelevich, M. and Sudakov, B. (1998) Finding a large hidden clique in a random graph. *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, 594–598.

Ames, B. P. W. and Vavasis, S. A. (2011) Nuclear norm minimization for the planted clique and biclique problems. *Math. Program.*, **129**, 69–89.

Amini, A. A. and Wainwright, M. J. (2009) High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Statist.*, **37**, 2877–2921.

Anandkumar, A., Ge. R., Hsu, D., Kakade, S. M. and Telgarsky, M. (2014) Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.*, **15**, 2773–2832.

Applebaum, B., Barak, B. and Wigderson, A. (2010) Public-key cryptography from different assumptions. *Proceedings of the Forty-Second ACM Symposium on Theory of Computing*, 171–180.

Arias-Castro, E., Verzelen, N. (2013) Community Detection in Dense Random Networks. *Ann. Statist.*, **42**, 940-969

Arora, S. and Barak, B. (2009) *Computational Complexity: A Modern Approach.* Cambridge University Press, Cambridge.

Aston, J. A. D. and Kirch, C. (2012) Evaluating stationarity via change-point alternatives with applications to fMRI data. *Ann. Appl. Stat.*, **6**, 1906–1948.

Aston, J. A. D. and Kirch, C. (2014) Change points in high dimensional settings. *arXiv preprint*, arxiv:1409.1771.

Aue, A., Hörmann, S., Horváth, L. and Reimherr, M. (2009) Break detection in the covariance structure of multivariate time series models. *Ann. Statist.*, **37**, 4046–4087.

Awasthi, P., Charikar, M., Lai, K. A. and Risteki, A. (2015) Label optimal regret bounds for online local learning. *J. Mach. Learn. Res. (COLT)*, **40**.

Bach, F., Ahipaşaoğlu, S. D., d'Aspremont, A. (2010) Convex relaxations for subset selection. *arXiv preprint*, arxiv:1006.3601.

Bai, J. (2010) Common breaks in means and variances for panel data. *J. Econometrics*, **157**, 78–92.

Baik, J., Ben Arous, G. and Péché, S. (2005) Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.*, **33**, 1643–1697.

Baik, J. and Silverstein, J. W. (2006) Eigenvalues of large sample covariance matrices of spiked population models. *J. Multivariate Anal.*, **97**, 1382–1408.

Bandeira, A. S., Dobriban, E., Mixon, D. G. and Sawin, W. F. (2012) Certifying the restricted isometry property is hard. *IEEE Trans. Inform. Theory*, **59**, 3448–3450.

Bandeira, A. S., Mixon, D. G. and Moreira, J. (2014) A conditional construction of restricted isometries. *International Mathematics Research Notices*, rnv385.

Baraniuk, R., Davenport, M., DeVore, R. and Wakin, M. (2008) A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, **28**, 253–263.

Berthet, Q. (2015) Optimal testing for planted satisfiability problems. *Electron. J. Statist.*, **9**, 298–317.

Berthet, Q. and Ellenberg, J. S. (2015) Detection of Planted Solutions for Flat Satisfiability Problems. *arXiv preprint*, arxiv:1502.06144.

Berthet, Q. and Rigollet P. (2013a) Optimal detection of sparse principal components in high dimension. *Ann. Statist.*, **41**, 1780–1815.

Berthet, Q. and Rigollet P. (2013b) Complexity theoretic lower bounds for sparse principal component detection. *J. Mach. Learn. Res. W&CP*, **30**, 1046–1066.

Bhaskara, A., Charikar, M., Chlamtac, E., Feige, U. and Vijayaraghavan, A. (2010) Detecting High Log-Densities an $O(n^{1/4})$ Approximation for Densest k-Subgraph. *Proceedings of the Forty-Second ACM symposium on Theory of Computing*, 201–210.

Birnbaum, A., Johnstone , I. M., Nadler , B. and Paul, D. (2013) Minimax bounds for sparse PCA with noisy high-dimensional data. *Ann. Statist.*, **41**, 1055–1084.

Blum, A., Kalai, A. and Wasserman, H. (2003) Noise-tolerant learning, the parity problem, and the statistical query model. *J. ACM*, **50**, 506–519.

Blumensath, T. and Davies, M. E. (2009) Iterative hard thresholding for compressed sensing. *Appl. Comput. Harmon. Anal.*, **27**, 265–274.

Bourgain, J., Dilworth, S., Ford, K. and Konyagin, S. (2011) Explicit constructions of RIP matrices and related problems. *Duke Math. J.*, **159**, 145–185.

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2011) Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, **3**, 1–122.

Bücher, A., Kojadinovic, I. Rohmer, T. and Seger, J. (2014) Detecting changes in cross-sectional dependence in multivariate time series. *J. Multivariate Anal.*, **132**, 111–128.

Cai, T. T., Ma, Z. and Wu, Y. (2013) Sparse PCA: Optimal rates and adaptive estimation. *Ann. Statist.*, **41**, 3074–3110.

Cai, T. T., Zhang, C.-H. and Zhou, H. H. (2010) Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.*, **38**, 2118–2144.

Candès, E. J. (2008) The restricted isometry property and its implications for compressed sensing. *C. R. Math. Acad. Sci. Paris*, **346**, 589–592.

Candès, E. J., Li, X., Ma, Y. and Wright, J. (2009) Robust Principal Component Analysis? *J. ACM* **58**, 1–37.

Candès E. J. and Recht, B. (2009) Exact matrix completion via convex optimization. *Found. of Comput. Math.*, **9**, 717–772.

Candès, E. J., Romberg, J. K. and Tao, T. (2006a) Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, **52**, 489–509.

Candès, E. J., Romberg, J. K. and Tao, T. (2006b) Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, **59**, 2006.

Candès E. J. and Tao, T. (2005) Decoding by Linear Programming. *IEEE Trans. Inform. Theory*, **51**, 4203–4215.

Candès E. J. and Tao, T. (2005) Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Inform. Theory*, **52**, 489–509.

Chan, Y.-b. and Hall, P. (2010) Using evidence of mixed populations to select variables for clustering very high-dimensional data. *J. Amer. Statist. Assoc.*, **105**, 798–809.

Chandrasekaran, V. and Jordan, M. I. (2013) Computational and statistical tradeoffs via convex relaxation. *Proc. Nat. Acad. Sci.*, **110**, E1181–E1190.

Chen, J. and Gupta, A. K. (1997) Testing and locating variance changepoints with application to stock prices. *J. Amer. Statist. Assoc.*, **92**, 739–747.

Chen, Y. and Xu, J. (2016) Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *J. Mach. Learn. Res.*, **17**, 1–57.

Chen, Y. and Ye, X. (2011) Projection onto a simplex. *arXiv preprint*, arxiv:1101.6081.

Cho, H. (2016) Change-point detection in panel data via double CUSUM statistic. *Electron. J. Stat.*, to appear.

Cho, H. and Fryzlewicz, P. (2015) Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *J. Roy. Statist. Soc. Ser. B*, **77**, 475–507.

Chun, H. and Sündüz, K. (2009) Expression quantitative trait loci mapping with multivariate sparse partial least squares regression. *Genetics*, **182**, 79–90.

Church, A. (1936) An unsolvable problem of elementary number theory. *Amer. J. Math.*, **58**, 345–363.

Cribben, I. and Yu, Y. (2016) Estimating whole brain dynamics using spectral clustering. *J. Roy. Statist. Soc. Ser. C*, to appear.

Csörgö, M. and Horváth, L. (1997) *Limit Theorems in Change-Point Analysis*. John Wiley and Sons, New York.

Cule, M., Samworth, R. J. and Stewart, M. (2010) Maximum likelihood estimation of a multi-dimensional log-concave density. *J. Roy. Statist. Soc. Ser. B. (with discussion)*, **72**, 545–607.

Darling, D. A. and Erdős, P. (1956) A limit theorem for the maximum of normalised sums of independent random variables. *Duke Math. J.*, **23**, 143–155.

d'Aspremont, A., Bach, F. and El Ghaoui, L. (2008) Optimal solutions for sparse principal component analysis. *J. Mach. Learn. Res.*, **9**, 1269–1294.

d'Aspremont, A. and El Ghaoui, L. (2011) Testing the nullspace property using semidefinite programming. *Math. Program.*, **127**, 123–144.

d'Aspremont, A. El Ghaoui, L., Jordan, M. I. and Lanckriet, G. R. G. (2007) A direct formulation for sparse PCA using semidefinite programming. *SIAM Rev.*, **49**, 434–448.

Dai, W. and Milenkovic, O. (2009) Subspace pursuit for compressive sensing signal reconstruction. *IEEE Trans. Inform. Theory*, **55**, 2230–2249.

Davis, C. and Kahan, W. M. (1970) The rotation of eigenvectors by a pertubation. III. *SIAM J. Numer. Anal.*, **7**, 1–46.

Deshpande, Y. and Montanari, A. (2014) Sparse PCA via covariance thresholding. *Adv. Neur. Inf. Proc. Sys.*, **27**, 334–342.

Diaconis, P. and Freedman D. (1980) Finite exchangeable sequences. *Ann. Probab.*, **8**, 745–764.

Donath, W. E. and Hoffman, A. J. (1973) Lower bounds for the partitioning of graphs. *IBM Journal of Research and Development* **17**, 420–425.

Donoho, D. L. (2006) Compressed sensing. *IEEE Trans. Inform. Theory*, **52**, 1289–1306.

Donoho, D. L., and Elad, M. (2003) mally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proc. Natl. Acad. Sci.*, **100**, 2197–2202.

Donoho, D. L., Elad, M. and Temlyakov, V. N. (2006) Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory*, **52**, 6–18.

Dümbgen, L. and Rufibach, K. (2009) Maximum likelihood estimation of a log-concave density and its distribution function: basic properties and uniform consistency. *Bernoulli*, **15**, 40–68.

Eldar, Y. C. and Kutyniok, G. (2012) *Compressed Sensing: Theory and Applications*. Cambridge University Press, Cambridge.

Enikeeva, F. and Harchaoui, Z. (2014) High-dimensional change-point detection with sparse alternatives. *arXiv preprint*, arxiv:1312.1900v2.

Fan, K. (1953) Minimax theorems. *Proc. Natl. Acad. Sci.*, **39**, 42–47.

Fan, J. and Han, X. (2013) Estimation of false discovery proportion with unknown dependence. *arXiv preprint*, arXiv:1305.7007.

Fan, J., Liao, Y. and Mincheva, M. (2013) Large covariance estimation by thresholding principal orthogonal complements. *J. Roy. Statist. Soc., Ser. B* **75**, 603–680.

Feige, U. Relations between average case complexity and approximation complexity. *Proceedings of the Thirty-Fourth Annual ACM Symposium on Theory of Computing*, 534–543.

Feige, U. and Krauthgamer, R. (2000) Finding and certifying a large hidden clique in a semirandom graph. *Random Structures Algorithms*, **16**, 195–208.

Feige, U. and Krauthgamer, R. (2003) The probable value of the Lovàsz–Schrijver relaxations for a maximum independent set. *SIAM J. Comput.*, **32**, 345–370.

Feige, U. and Ron, D. (2010) Finding hidden cliques in linear time. *Discrete Math. Theor. Comput. Sci. Proc.*, 189–204.

Feldman, V., Grigorescu, E., Reyzin, L., Vempala, S. S. and Xiao, Y. (2013) Statistical algorithms and a lower bound for detecting planted cliques. *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, 655–664.

Feldman, V., Perkins, W. and Vempala, S. (2015) On the Complexity of Random Satisfiability Problems with Planted Solutions. *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, 77–86.

Frick, K., Munk, A. and Sieling, H. (2014) Multiscale change point inference. *J. Roy. Statist. Soc. Ser. B*, **76**, 495–580.

Fryzlewicz, P. (2014) Wild binary segmentation for multiple change-point detection. *Ann. Statist.*, **42**, 2243–2281.

Gabay, D. and Mercier, B. (1976) A dual algorithm for the solution of nonlinear variational problems via finite element approximations. *Comput. Math. Appl.*, **2**, 17–40.

Gao, C., Ma, Z. and Zhou, H. H. (2014) Sparse CCA: Adaptive estimation and computational barriers. *arXiv preprint*, arxiv:1409.8565.

Allen, G. I., and Maletić-Savatić, M. (2011) Sparse non-negative generalized PCA with applications to metabolomics. *Bioinformatics* **27**, 3029–3035.

Golub, G. H. and Van Loan, C. F. (1996) *Matrix Computations*. The Johns Hopkins University Press, Baltimore, Maryland.

Grimmett, G. R. and McDiarmid C. J. H. (1975) On colouring random graphs. *Math. Proc. Cambridge Philos. Soc.*, **77**, 313–324.

Hajek, B., Wu, Y. and Xu, J. (2014) Computational lower bounds for community detection on random graphs. *Proceedings of the Twenty-Eighth Conference on Learning Theory*, 899–928.

Hampel, F. R. (1974) The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.*, **69**, 383–393.

Hazan, E. and Krauthgamer, R. (2011) How hard is it to approximate the best Nash equilibrium? *SIAM J. Comput.*, **40**, 79–91.

Henry, D., Simani, S. and Patton, R. J. (2010) Fault detection and diagnosis for aeronautic and aerospace missions. In Edwards, C., Lombaerts, T., and Smaili H., eds, *Fault Tolerant Flight Control — A Benchmark Challenge*, 91–128. Springer-Verlag, Berlin.

Horn, R. A. and Johnson, C. R. (2012) *Matrix Analysis*. Cambridge University Press.

Horváth, L., Kokoszka, P. and Steinebach, J. (1999) Testing for changes in dependent observations with an application to temperature changes. *J. Multivariate Anal.*, **68**, 96–199.

Horváth, L. and Rice, G. (2014) Extensions of some classical methods in change point analysis. *TEST*, **23**, 219–255.

Horváth, L. and Hušková, M. (2012) Change-point detection in panel data. *J. Time Ser. Anal.*, **33**, 631–648.

Hubert, L. and Arabie, P. (1985) Comparing partitions. *J. Classification*, **2**, 193–218.

Jerrum, M. (1992) Large cliques elude the Metropolis process. *Random Structures Algorithms*, **3**, 347–359.

Jirak, M. (2015) Uniform change point tests in high dimension. *Ann. Statist.*, **43**, 2451–2483.

Johnstone, I. M. and Lu, A. Y. (2009) On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.*, **104**, 682–693.

Jolliffe, I. T., Trendafilov, N. T. and Uddin, M. (2003) A modified principal component technique based on the LASSO, *J. Comput. Graph. Statist.*, **12**, 531–547.

Journée, M., Nesterov, Y., Richtárik, P. and Sepulchre, R. (2010) Generalized power method for sparse principal component analysis. *J. Mach. Learn. Res.*, **11**, 517–553.

Juditsky, A. and Nemirovski, A. (2011) On verifiable sufficient conditions for sparse signal recovery via ℓ_1 minimization. *Math. Program.*, **127**, 57–88.

Juels, A. and Peinado, M. (2000) Hiding cliques for cryptographic security. *Des. Codes Cryptogr.*, **20**, 269–280.

Karp, R. M. (1972) Reducibility among combinatorial problems. In Miller, R. et al., eds., *Complexity of Computer Computations*, 85–103. Springer, New York.

Kirch, C., Mushal, B. and Ombao, H. (2015) Detection of changes in multivariate time series with applications to EEG data. *J. Amer. Statist. Assoc.*, **110**, 1197–1216.

Killick, R., Fearnhead, P. and Eckley, I. A. (2012) Optimal detection of changepoints with a linear computational cost. *J. Amer. Stat. Assoc.*, **107**, 1590–1598.

Kim, A. K.-H. and Samworth R. J. (2016) Global rates of convergence in log-concave density estimation. *Ann. Statist.*, to appear.

Koiran, P. and Zouzias, A. (2012) Hidden cliques and the certification of the restricted isometry property. *IEEE Trans. Inform. Theory*, **60**, 4999–5006.

Kučera, L. (1995) Expected complexity of graph partitioning problems. *Discrete Appl. Math.*, **57**, 193–212.

Kukush, A., Markovsky, I. and Van Huffel, S. (2002) Consistent fundamental matrix estimation in a quadratic measurement error model arising in motion analysis. *Comput. Stat. Data An.*, **41**, 3–18.

Lanczos, C. (1950) An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Res. Nat. Bur. Stand.*, **45**, 255–282.

Laurent, B. and Massart, P. (2000) Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, **28**, 1302–1338.

Lavielle, M. and Teyssiere, G. (2006) Detection of multiple change-points in multivariate time series. *Lithuanian Math. J.*, **46**, 287–306.

Lee, K. and Bresler, Y. (2008) Computing performance guarantees for compressed sensing. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 5129–5132.

Ma, Z. (2013) Sparse principal component analysis and iterative thresholding. *Ann. Statist.*, **41**, 772–801.

Ma, Z. and Wu, Y. (2015) Computational barriers in minimax submatrix detection. *Ann. Statist.*, **43**, 1089–1116.

Mallat, S. (1999) *A Wavelet Tour of Signal Processing*. Academic Press, Cambridge, MA.

Majumdar, A. (2009) Image compression by sparse PCA coding in curvelet domain. *Signal, Image and Video Processing*, **3**, 27–34.

Massart, P. (2007) *Concentration Inequalities and Model Selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII–2003*. Springer, Berlin/Heidelberg.

McDiarmid, C. (1989) On the method of bounded differences. *Surveys in Combinatorics*, **141**, 148–188.

Naikal, N., Yang A. Y. and Sastry S. S. (2011) Informative feature selection for object recognition via sparse PCA. *IEEE International Conference on Computer Vision (ICCV)*, 818–825.

Needell, D. and Tropp, J. A. (2009) CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Appl. Comput. Harmon. Anal.*, **26**, 301–321.

Nemirovski, A. (2004) Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM J. Optim.*, **15**, 229–251.

Nesterov, Yu. (2005) Smooth minimization of nonsmooth functions. *Math. Program., Ser. A*, **103**, 127–152.

Olshen, A. B., Venkatraman, E. S., Lucito, R. and Wigler, M. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biometrika*, **5**, 557–572.

Ombao, H., Von Sachs, R. and Guo, W. (2005) SLEX analysis of multivariate nonstationary time series. *J. Amer. Statist. Assoc.*, **100**, 519–531.

Page, E. S. (1955) A test for a change in a parameter occurring at an unknown point. *Biometrika*, **42**, 523–527.

Parkhomenko, E., Tritchler, D., and Beyene, J. (2009) Sparse canonical correlation analysis with application to genomic data integration. *Stat. Appl. Genet. Mol. Biol.*, **8**, 1–34.

Paul, D. (2007) Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica*, **17**, 1617–1642.

Peng, T., Leckie, C. and Ramamohanarao, K. (2004) Proactively detecting distributed denial ofservice attacks using source IP address monitoring. In Mitrou, N., Kontovasilis, K., Rouskas, G. N., Iliadis, I. and Merakos, L. eds, *Networking 2004*, 771–782. Springer-Verlag, Berlin.

Preuß, P., Puchstein, R. and Dette, H. (2015) Detection of multiple structural breaks in multivariate time series. *J. Amer. Statist. Assoc.*, **110**, 654–668.

Rand, W. M. (1971) Objective criteria for the evaluation of clustering methods. *J. Amer. Statist. Assoc.*, **66**, 846–850.

Rauhut, H. and Foucart, S. (2013) *A Mathematical Introduction to Compressive Sensing*. Birkhäuser, Basel.

Rohe, K., Chatterjee, S. and Yu, B. (2011) Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.*, **39**, 1878–1915.

Shabalin, A. A. and Nobel, A. B. (2013) Reconstruction of a low-rank matrix in the presence of Gaussian noise. *J. Multivariate Anal.*, **118**, 67–76.

Shen, D., Shen, H. and Marron, J. S. (2013) Consistency of sparse PCA in high dimension, low sample size contexts. *J. Multivariate Anal.*, **115**, 317–333.

Shepp, L. A. (1971) First passage time for a particular Gaussian process. *Ann. Math. Statist.*, **42**, 946–951.

Shorack, G. R. and Wellner, J. A. (1986) *Empirical Processes with Applications to Statistics*. Wiley, New York.

Slepian, D. (1961) First passage time for a particular Gaussian process. *Ann. Math. Statist.*, **32**, 610–612.

Slepian, D. (1962) The one-sided barrier problem for Gaussian noise. *Bell System Techinical Journal*, **41**, 463–501.

Sparks, R., Keighley, T. and Muscatello, D. (2010) Early warning CUSUM plans for surveillance of negative binomial daily disease counts. *J. Appl. Stat.*, **37**, 1911–1930.

Stewart, G. W. and Sun, J.-G. (1990) *Matrix Perturbation Theory*. Academic Press, Inc.: San Diego, California.

Sun T. and Zhang, C.-H. (2012) Calibrated elastic regularization in matrix completion. *Adv. Neur. Inf. Proc. Sys.*, **25**, Ed. F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger. MIT Press: Cambridge, Massachusetts.

Tan, K. M., Petersen, A. and Witten, D. (2014) Classification of RNA-seq data. In S. Datta and D. Nettleton, eds. *Statistical Analysis of Next Generation Sequencing Data*, 219–246. Springer, New York.

Tillmann, A. N. and Pfetsch M. E. (2014) The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Trans. Inform. Theory*, **60**, 1248–1259.

Turing, A. (1936) On computable numbers, with an application to the Entscheidungsproblem. *Proc. London Math. Soc.*, **2**, 230–265.

van de Geer, S. (2000) *Empirical Processes in M-Estimation*. Cambridge University Press, Cambridge.

van der Vaart, A. W. and Wellner, J. A. (1996) *Weak Convergence and Empirical Processes*. Springer, New York.

Van Huffel, S. and Vandewalle, J. (1989) On the accuracy of total least squares and least squares techniques in the presence of errors on all data. *Automatica* **25**, 765–769.

Venkatraman, E. S. (1992) Consistency results in multiple change-point problems. Doctoral dissertation, to the Department of Statistics, Stanford University.

Vershynin, R. (2012) Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok, eds. *Compressed Sensing, Theory and Applications*, 210–268. Cambridge University Press, Cambridge.

von Luxburg, U. (2007) A tutorial on spectral clustering. *Statist. Comput.*, **17**, 395–416.

Vu, V. Q., Cho, J., Lei, J. and Rohe, K. (2013) Fantope projection and selection: a near-optimal convex relaxation of sparse PCA. *Adv. Neur. Inf. Proc. Sys.*, **26**, 2670–2678.

Vu, V. Q. and Lei, J. (2013) Minimax sparse principal subspace estimation in high dimensions. *Ann. Statist.*, **41**, 2905–2947.

Wang, T., Berthet, Q. and Plan, Y. (2016) Average-case hardness of RIP certification. *arXiv preprint*, arxiv:1605.09646.

Wang, T., Berthet, Q. and Samworth, R. J. (2016) Statistical and computational trade-offs in Estimation of Sparse Pincipal Components. *Ann. Statist.*, to appear.

Wang, T. and Samworth, R. J. (2016a) High-dimensional changepoint estimation via sparse projection. *arXiv preprint*, arxiv:1606.06246.

Wang, T. and Samworth, R. J. (2016b) InspectChangepoint: high-dimensional changepoint estimation via sparse projection. R package version 1.0, https://cran.r-project.org/web/packages/InspectChangepoint/.

Wang, D., Lu, H.-C. and Yang, M.-H. (2013) Online object tracking with sparse prototypes. *IEEE Trans. Image Process.*, **22**, 314–325.

Wang, Z., Lu, H. and Liu, H. (2014) Tighten after Relax: Minimax-Optimal Sparse PCA in Polynomial Time. *Adv. Neur. Inf. Proc. Sys.*, **27**, 3383–3391.

Wedin, P.-Å. (1972) Perturbation bounds in connection with singular value decomposition. *BIT* **12**, 99–111.

Weyl, H. (1912) Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf der Theorie der Hohlraumstrahlung). *Math. Ann.*, **71**, 441–479.

Wilkinson, J. H. (1965) *The Algebraic Eigenvalue Problem*. Clarendon Press, Oxford.

Witten, D. M., Tibshirani, R. and Hastie, T. (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, **10**, 515–534.

Yu, B. (1997) Assouad, Fano and Le Cam. In Pollard, D., Torgersen, E. and Yang G. L., eds., *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, 423–435. Springer, New York.

Yu, Y., Wang, T. and Samworth, R. J. (2015) A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, **102**, 315–323.

Yuan, X.-T. and Zhang, T. (2013) Truncated power method for sparse eigenvalue problems. *J. Mach. Learn. Res.*, **14**, 899–925.

Zhang, N. R., Siegmund, D. O., Ji, H. and Li, J. Z. (2010) Detecting simultaneous changepoints in multiple sequences. *Biometrika*, **97**, 631–645.

Zhang, Y., Wainwright, M. J. and Jordan, M. I. (2014) Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. *J. Mach. Learn. Res. W&CP*, **35**, 921–948.

Zou, H., Hastie, T. and Tibshirani, R. (2006) Sparse principal components analysis. *J. Comput. Graph. Statist.*, **15**, 265–286.